The fourth radiation transfer model intercomparison (RAMI-IV): Proficiency testing of canopy reflectance models with ISO-13528

J.-L. Widlowski,[1] B. Pinty,[1] M. Lopatka,[1] C. Atzberger,[2] D. Buzica,[3] M. Chelle,[4] M. Disney,[5,6] J-P. Gastellu-Etchegorry,[7] M. Gerboles,[1] N. Gobron,[1] E. Grau,[7] H. Huang,[8] A. Kallel,[9] H. Kobayashi,[10,11] P. E. Lewis,[5,6] W. Qin,[12] M. Schlerf,[13] J. Stuckens,[14] and D. Xie[15]

Received 12 December 2012; revised 11 May 2013; accepted 13 May 2013; published 1 July 2013.

[1] The radiation transfer model intercomparison (RAMI) activity aims at assessing the reliability of physics-based radiative transfer (RT) models under controlled experimental conditions. RAMI focuses on computer simulation models that mimic the interactions of radiation with plant canopies. These models are increasingly used in the development of satellite retrieval algorithms for terrestrial essential climate variables (ECVs). Rather than applying ad hoc performance metrics, RAMI-IV makes use of existing ISO standards to enhance the rigor of its protocols evaluating the quality of RT models. ISO-13528 was developed "to determine the performance of individual laboratories for specific tests or measurements." More specifically, it aims to guarantee that measurement results fall within specified tolerance criteria from a known reference. Of particular interest to RAMI is that ISO-13528 provides guidelines for comparisons where the true value of the target quantity is unknown. In those cases, "truth" must be replaced by a reliable "conventional reference value" to enable absolute performance tests. This contribution will show, for the first time, how the ISO-13528 standard developed by the chemical and physical measurement communities can be applied to proficiency testing of computer simulation models. Step by step, the pre-screening of data, the identification of reference solutions, and the choice of proficiency statistics will be discussed and illustrated with simulation results from the RAMI-IV "abstract canopy" scenarios. Detailed performance statistics of the participating RT models will be provided and the role of the accuracy of the reference solutions as well as the choice of the tolerance criteria will be highlighted.

Citation: Widlowski, J.-L., et al. (2013), The fourth radiation transfer model intercomparison (RAMI-IV): Proficiency testing of canopy reflectance models with ISO-13528, *J. Geophys. Res. Atmos.*, *118*, 6869–6890, doi:10.1002/jgrd.50497.

1. Introduction

[2] Physics-based radiative transfer (RT) models simulate the interactions of solar radiation within a given medium (e.g., clouds, plant canopies, etc.). Increasingly, these models contribute to the quantitative interpretation of remote sensing observations. Model simulations, for example, can be used to generate look-up-tables, to train neural networks, or to develop parametric formulations that are then embedded in quantitative retrieval algorithms. A case in point are many of the current LAI, FAPAR, and surface albedo products that are derived from global medium-resolution sensors. The quality of RT models is thus essential if accurate and reliable information are to be derived from Earth Observation (EO) data. In fact, the reliability of model simulations

[1] Institute for Environment and Sustainability, DG Joint Research Centre, European Commission, Ispra, Italy.

[2] Institute of Surveying, Remote Sensing and Land Information, University of Natural Resources and Life Sciences, Vienna, Austria.

[3] Industrial Emissions, Air Quality and Noise Unit, DG Environment, European Commission, Brussels, Belgium.

[4] Institut National de la Recherche Agronomique, Thiverval-Grignon, France.

[5] Department of Geography, University College London, UK.

[6] Department of Meteorology, NERC National Centre for Earth Observation (NCEO), University of Reading, Reading, UK.

[7] Centre d'Etudes Spatiales de la BIOsphère, Toulouse, France.

[8] Beijing Forestry University, Beijing, China.

[9] Institut Supérieur d'Electronique et de Communication de Sfax, Tunisia.

[10] Department of Environmental Science, Policy and Management, University of California, Berkeley, California, USA.

[11] Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan.

[12] Science Systems and Applications, Inc., Greenbelt, Maryland, USA.

[13] Département Environnement et Agro-biotechnologie, Centre de Recherche Public - Gabriel Lippmann, Belvaux, Luxembourg.

[14] Biosystems Department, Katholieke Universiteit Leuven, Leuven, Belgium.

[15] Research Center for Remote Sensing and GIS, School of Geography, Beijing Normal University, Beijing, China.

Corresponding author: J.-L. Widlowski, Institute for Environment and Sustainability, European Commission's DG Joint Research Centre, TP272, I-21027 Ispra (VA), Italy. (Jean-Luc.Widlowski@jrc.ec.europa.eu)

2169-897X/13/10.1002/jgrd.50497

should at least be comparable to the space sensor uncertainties documented by vicarious calibration efforts. This is particularly so in the context of climate studies where both accuracy and stability requirements for satellite-derived essential climate variables (ECVs) are increasingly stringent [*GCOS*, 2011].

[3] Obtaining accurate satellite and in situ estimates of terrestrial ECVs is highly challenging. Most field validation efforts of quantitative EO products over land are still in a pre-standardized state. As such, it is not surprising that neither funding agencies nor environmental legislation currently enforces absolute quality criteria on satellite-derived quantitative surface information. The situation is rather different, however, when it comes to laboratory or in situ measurements in the field of environmental chemistry. Here a large body of legislation exists at both national and supranational level that (1) defines acceptable concentration ranges and limits of target substances, (2) regulates the manner in which these quantities should be measured and analysed, and (3) indicates procedures to deal with eventual exceedances. Fundamental to such a framework are both the availability of error-characterized reference methodologies and the existence of standardized procedures allowing evaluation of the quality of alternative measurement techniques. Of particular interest in this context is the formulation of methodological standards that allow for regular testing of the *proficiency* of laboratories. The goal of such procedures is to guarantee that the results—obtained by performing comparable analyses with laboratory-specific measurement methods—fall within specified tolerance criteria from a known reference [*Hund et al.*, 2000].

[4] Currently, the usage of these community-approved and internationally applied quality assurance standards is limited to the evaluation of chemical and physical measurement procedures, e.g., *Gerboles et al.* [2011]. If it were possible to apply these standards to the comparison of physics-based computer simulation models (and subsequently also satellite retrieval algorithms), then the rigorousness of such comparison efforts would certainly benefit, their findings would become more authoritative, and the acceptance of their outcome would be broader. Such a transfer of context appears feasible since the crux of absolute verification schemes remains essentially the same irrespective of whether one deals with measurements (i.e., laboratory or in situ data) or simulations (i.e., model or algorithm outputs). In both scenarios, the true value of the target quantity is generally unknown, and thus, it is foremost the definition of a reliable "conventional reference value" [*JCGM*, 2008] that is required to carry out absolute performance tests. At the same time, however, appropriate tolerance criteria must be defined and suitable statistical tools selected to determine whether a given bias is acceptable or not. In this contribution, the international standard *ISO 13528* [2005] (and in part also *ISO 5725-2* [1994]) is applied to the data submitted to the fourth phase of the RAdiation transfer Model Intercomparison (RAMI) exercise.

[5] As an open, self-organizing activity of the canopy RT modeling community the RAMI exercise has focused, since 1999, on the evaluation of models simulating bidirectional reflectance factors (BRFs) and radiative fluxes for 1-D and 3-D vegetation canopies [*Pinty et al.*, 2001, 2004]. The first three phases of RAMI, which concentrated on relatively simple and often abstracted plant environments, allowed participants to identify coding errors and to improve some of the RT formulations in their models. As such, model agreement increased and reached ~1% on average among the 3-D Monte Carlo models participating in the third phase of RAMI [*Widlowski et al.*, 2007b]. This in turn enabled the definition of a "surrogate truth" data set and the subsequent development of a web-based benchmarking facility known as the RAMI On-line Model Checker (ROMC) [*Widlowski et al.*, 2007a]. "Credible" Monte Carlo models from RAMI-3 have also been used to evaluate the quality of RT formulations embedded in the land surface schemes of soil-vegetation-atmosphere transfer (SVAT), numerical weather prediction (NWP), and global circulation models [*Widlowski et al.*, 2011]. With these achievements, the scope of RAMI was ready to be expanded toward more complex and realistic representations of plant environments as well as the simulation of new types of measurements and remote sensing devices.

[6] This paper is subdivided as follows: Section 2 summarizes the experimental setup and measurement definitions used by the fourth phase of RAMI (RAMI-IV). Section 3 presents the outcome of several consistency checks that were applied to screen the contributions of the participating models. Section 4 provides an overview of ISO-13528 and shows how this can be applied to define consensus reference values for the RAMI-IV test cases. In section 5, the performance of the participating models will be presented. Section 6 concludes with a series of observations and remarks.

2. The Fourth Phase of RAMI

[7] In February 2009, potential participants were invited by email to contribute to RAMI-IV. A dedicated website (http://rami-benchmark.jrc.ec.europa.eu/) had been set up containing detailed descriptions of the prescribed test cases. Instructions were also provided as to what radiative quantities had to be simulated. The task of the participants then consisted in (1) representing the prescribed canopy architectures within their respective RT model(s), (2) executing their models to simulate the prescribed radiation quantities under predefined illumination and viewing conditions, and (3) formatting and uploading the output of their models according to the RAMI specifications. Similar to previous phases of RAMI the collection of results, their analysis, and any eventual feedbacks to the model operators were carried out by the European Commission's Joint Research Centre (JRC) in Ispra, Italy. Table 1 lists the models and operators that submitted simulations for the abstract canopy experiments of RAMI-IV.

2.1. RAMI-IV Abstract Canopies

[8] Although RAMI-IV consisted of two separate sets of experiments only those pertaining to the "abstract canopy" category will be used in this work. Figure 1 provides a graphical overview of the prescribed architectural scenarios. As can be seen, the test cases were based exclusively on finite-sized disc-shaped scatterers (i.e., leaves) that were characterized by Lambertian scattering properties and various orientation distributions [*Goel and Strebel*, 1984]. The scatterers were confined to spherical, cylindrical, or slab-like volumes floating above a flat background. The position

Table 1. List of RT Models, Associated Publications and Operators Contributing to the “Abstract Canopy” Category of RAMI-IV[a]

Model Name	Model Reference(s)	Operator	RAMI-3 Participant
1-D models:			
`FDM`	*Kallel* [2010, 2012]	A. Kallel	No
`1/2-discret`	*Gobron et al.* [1997]	N. Gobron	Yes
3-D models:			
`DART`	*Gastellu-Etchegorry et al.* [1996, 2004]	E. Grau and J-P. Gastellu	Yes[a]
`FLiES`[b]	*Kobayashi and Iwabuchi* [2008]	H. Kobayashi	No
`INFORM`	*Atzberger* [2000]; *Schlerf and Atzberger* [2006]	C. Atzberger and M. Schlerf	No
`librat`[b]	*Lewis* [1999]; *Disney et al.* [2009]	M. Disney and P. Lewis	Yes
`parcinopy`[b]	*Chelle* [1997, 2006]	M. Chelle	No
`pbrt`[b]	*Pharr and Humphreys* [2010]	J. Stuckens	No
`RaySpread`[b]	*Widlowski et al.* [2006]	J-L. Widlowski	Yes
`raytran`[b]	*Govaerts* [1995]	J-L. Widlowski	Yes
`RGM`	*Qin and Gerstl* [2000]	D. Xie	Yes
`RGM2`	*Liu et al.* [2007] and *Huang et al.* [2009]	H. Huang	No

[a]The `DART` model that participated in RAMI-3 was relying on voxels—having statistical scattering properties—to build individual tree crowns and canopies. The `DART` model contributing to RAMI-IV was a newer version capable of representing individual scatterers.

[b]This model makes use of Monte Carlo ray-tracing techniques.

and orientation of every single scatterer was made available to the participants. In all cases, the illumination conditions consisted only of a direct light source.

[9] Three different types of spatially homogeneous canopies were proposed in RAMI-IV, that is, (1) canopies with anisotropic backgrounds, (2) canopies composed of two horizontal layers having different spectral and structural properties, and (3) canopies composed of two adjacent sections having different spectral and structural properties. Similarly, the RAMI-IV heterogeneous experiments were subdivided into three architectural classes, that is, (1) canopies composed of scatterers confined to spherical volumes floating above an anisotropic background, (2) two layer canopies where a series of larger spherical volumes (containing scatterers) floated above an understorey composed of many smaller spherical volumes (also containing scatterers), and (3) canopies with scatterers confined to cylindrical volumes that were inclined at a fixed angle to the background. The BRFs of the anisotropically scattering backgrounds were modelled with the parametric RPV model [*Rahman et al.*, 1993b, 1993a]. The RPV parameters had been chosen such as to mimic the properties of snow, bare soil and understorey vegetation.

2.2. RAMI-IV Measurements

[10] For each RAMI-IV test case, a series of “measurements” had to be simulated under well-described illumination and observation conditions. Model simulations had to contain six significant digits. In addition to the measurement types listed below, simulations of the local

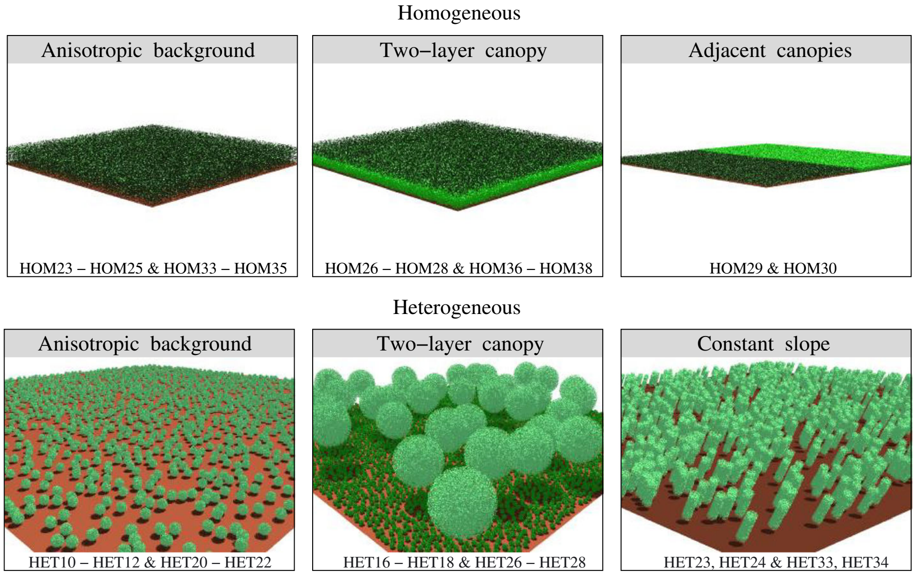

Figure 1. Graphical representations of the (top panels) homogeneous and (bottom panels) heterogeneous canopy architectures prescribed within the “abstract canopy” category of RAMI-IV. Experiment identifiers are indicated below the pictures. Colors are for visualization purposes only and do not reflect actual spectral properties.

Table 2. Overview of the Type of Measurements Performed by RT Models Contributing to the “Abstract Canopy” Cases of RAMI-IV[a]

Model Name	BRFs				Fluxes		Transmission					Lidar	
	tot	uc	co	mlt	DHR	fabs	tot	coco	uc	vprof	loc_dir	tot	sgl
FDM	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	-
1/2-discret	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	-
DART	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓
FLiES	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓
INFORM	✓	-	-	-	-	-	-	-	-	-	-	-	-
librat	✓	✓	✓	✓	-	-	-	✓	-	-	-	✓	✓
parcinopy	✓	-	-	✓	✓	✓	✓	-	✓	✓	-	-	-
pbrt	✓	✓	✓	✓	✓	✓[b]	✓	✓	✓	✓	-	-	-
rayspread	✓	✓	✓	✓	-	-	-	-	-	-	✓	-	-
raytran	-	-	-	-	✓	✓	✓	✓	✓	✓	-	✓	✓
RGM	✓	-	-	-	-	-	-	-	-	-	-	-	-
RGM2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	✓

[a] BRF stands for bidirectional reflectance factor, uc for radiation uncollided with the vegetation, co for radiation single-collided with vegetation, and mlt for multiple-collided radiation; tot is the contribution from all available radiation components, coco stands for radiation collided with the canopy only, sgl stands for radiation collided once with either vegetation or background, DHR stands for directional hemispherical reflectance, fabs stands for fraction of absorbed radiation, vprof stands for vertical transmission profiles, and loc_dir stands for direct transmission at a particular location within the canopy/scene.

[b] Reconstructed by imposing energy conservation ($\Delta F = 0$ in equation (1)) for test cases with Lambertian backgrounds.

uncollided transmission along a transect at the lower boundary had been asked for. These simulations were intended to mimic the radiative quantities that would be gathered by the TRAC instrument [*Chen and Cihlar*, 1995] if placed within the heterogeneous abstract canopies of RAMI-IV. However, only the rayspread model submitted this type of simulation results. Table 2 provides an overview of the contributions submitted by the various participating RT models.

2.2.1. Bi-Directional Reflectance Factors (BRFs)

[11] BRFs had to be generated for view zenith angles at 2° interval from ±1° to ±75° along the principal plane

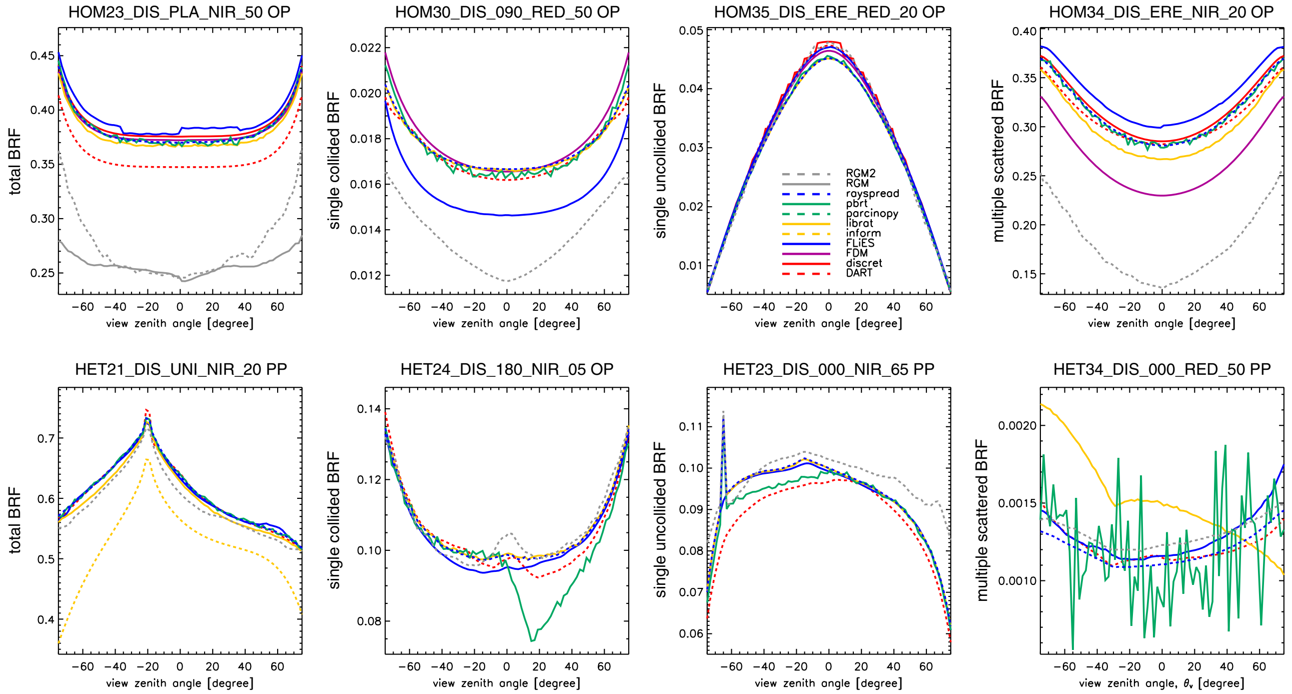

Figure 2. A selection of examples showing BRF simulations from RT models participating in the RAMI-IV “abstract canopy” test cases. The top row pertains to homogeneous canopies, whereas the bottom row relates to heterogeneous test cases. Going from left to right the panels relate to the total BRF, the single-collided by vegetation BRF, the uncollided by vegetation BRF, and the multiple-collided BRF. The labels above each graph are RAMI identifier tags where the first three groups of text describe the canopy target and the last three groups identify the spectral band (RED/NIR), the solar zenith angle (50/20/05/65), and the plane of observation (OP/PP), respectively.

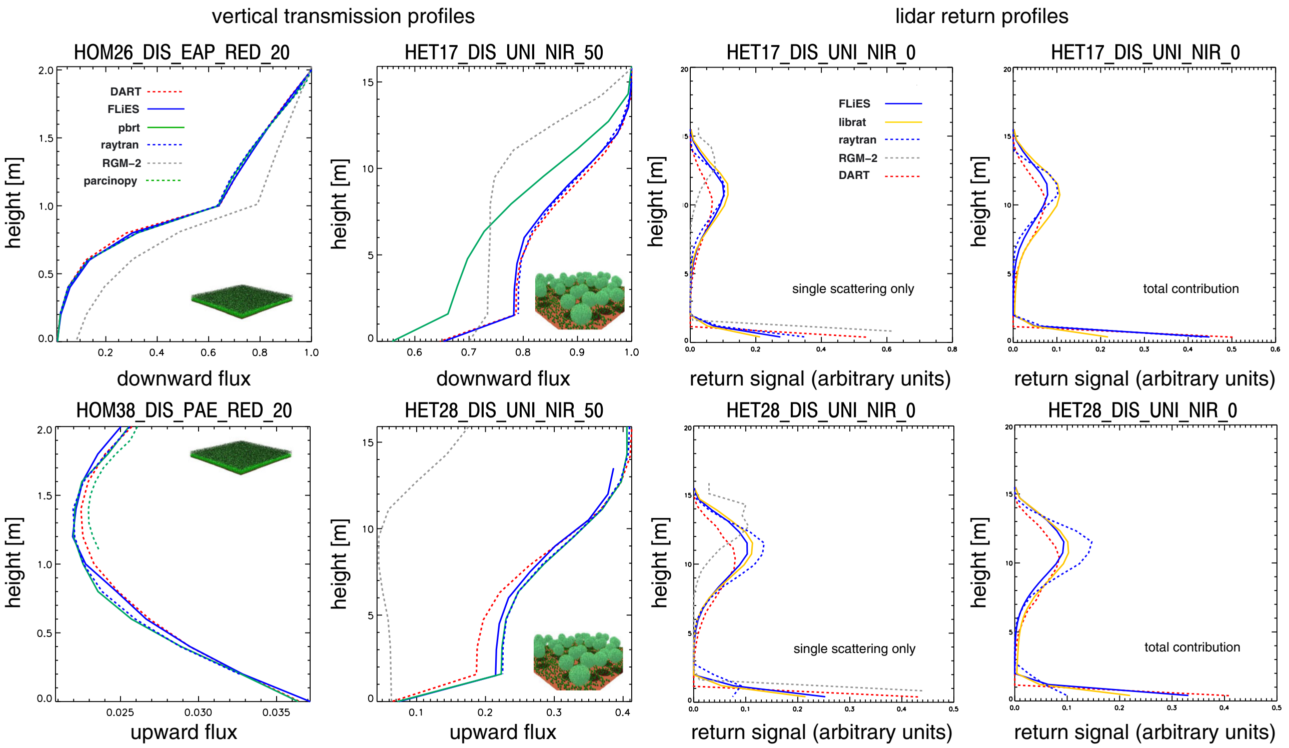

Figure 3. Left-hand panels show vertical transmission profiles selected from among the RAMI-IV "abstract canopy" test cases. Simulations were normalized with the incident flux at the top-of-canopy level. Upper panels show downward fluxes, while lower panels show upward fluxes in the red and NIR spectral regimes. Right-hand panels show lidar return profiles normalized such that their vertical integral equals unity. The labels above each graph are RAMI identifier tags where the first three groups of text describe the canopy target and the last two groups identify the spectral band (RED/NIR) and solar zenith angle (50/20/0), respectively.

(PP) and the orthogonal plane (OP). This setup allowed to avoid BRF simulations coinciding exactly with the peak of the hotspot (i.e., the retro-reflection direction). The BRF simulations in the PP and OP included the total BRF, the single-collided by the vegetation BRF component, the uncollided by the vegetation BRF component, and the multiple scattered BRF component. A total of 58,356 BRF files were received overall. If one removes the multiple/updated submissions, only 21,423 unique BRF files remained (containing 1,628,148 BRF values) that were included in the proficiency testing.

[12] Figure 2 provides examples of the variability of model simulated BRFs in the red and near-infrared (NIR) spectral domain along the PP or OP for both homogeneous (top panels) and heterogeneous (bottom panels) abstract canopies. In general, one will find a cluster of models having similar output and one (or more) models with somewhat different simulation results. However, without any supporting auxiliary information one should refrain from concluding that the clustering indicates proximity to the true value. Interestingly, the models that deviate from the clustering area are not always the same in Figure 2. In addition, the model deviations can occur across all or only a selection of the simulated viewing conditions. This highlights the need to evaluate RT models over a great number of test cases spanning a large set of architectural, spectral, and directional conditions.

2.2.2. Hemispherical Fluxes

[13] Prescribed hemispherical flux measurements included the directional hemispherical reflectance (DHR), the foliage absorption, and the total transmission at the lower boundary level. RAMI-IV also required simulations of the transmission components that reached the underlying background without undergoing any collisions within the canopy volume (ftran_uc_dir) or else that had at least one interaction with the canopy volume (ftran_coco_dir). Overall, the total number of submitted flux files was 31,218. In addition, the vertical profiles of total upward and downward transmissions through the canopy were required. Of the 5,869 files that were received with vertical transmission profiles (including erroneous submissions) only 2,023 files (or 66,759 data points) remained for the final analysis.

[14] The left half of Figure 3 shows a selection of vertical flux profile simulations (normalized by the incident flux at the top-of-canopy level) for several of the RAMI-IV abstract canopy test cases. The top (bottom) row shows downward (upward) fluxes, while the first (second) column relates to homogeneous (heterogeneous) canopy cases in the red (NIR) spectral domain. In general, the model simulations tend to be somewhat more clustered over the spatially homogeneous test cases.

2.2.3. Lidar

[15] RAMI-IV proposed to simulate the return signal of a waveform LIDAR instrument operating in the NIR. Both

the total return signal and that accounting for the first order of scattering in the canopy had to be generated. More specifically, the models should mimic an instantaneous pulse of radiation that resulted in a circular footprint of 50 m diameter at the top-of-canopy height level. Only photons exiting from this uniformly illuminated target area could actually contribute to the lidar return signal. The waveform signal itself was to be discretised into contributions originating from 20 height intervals/bins of equal thickness. The field of view (FOV) of the detector was set to 24 mrad, and its height was equal to 2000 m (above the background). The radius of the telescope collecting the returned photons was 1 m. The actual quantity to report was the amount of radiation that was scattered back up from a given height interval/bin into the field of view of the detector normalized by the incident radiation within the footprint area.

[16] The definition of the lidar measurements in RAMI-IV lead to different interpretations of the target quantities. As such, the results displayed in the right half of Figure 3 were normalized for better visual comparison in a manner such that the sum of contributions from all height intervals/bins equals unity. The rightmost panels show the total return signal, whereas the inner panels show the return signal after one scattering event. Although the exact lidar profiles are somewhat different, the peak of their overstory contribution occurs at very similar heights. This is especially the case for the `FLiES`, `librat`, and `raytran` models. Due to the differences in the formats of the received data set, it was decided not to pursue their analysis further at this stage but to refine the simulation requirements on the RAMI website first.

3. Model Consistency Checks

[17] In line with the requirements of ISO-13528 and also in analogy to previous phases of RAMI, a series of model consistency checks were carried out prior to the actual proficiency testing.

3.1. Energy Conservation

[18] Energy conservation describes the fact that all radiation entering or exiting a given plant canopy volume must be in balance with the amount of energy that is being absorbed by this volume. In RAMI-IV, energy conservation can only be evaluated for test cases having backgrounds with Lambertian scattering properties. In those cases, the deviation of model m from energy conservation for any particular structural (ζ), spectral (λ) and illumination (Ω_i) related conditions, can be defined as

$$\Delta\mathrm{F}_m(\lambda,\zeta,\Omega_\mathrm{i}) = 1 - [A_m(\lambda,\zeta,\Omega_\mathrm{i}) + R_m(\lambda,\zeta,\Omega_\mathrm{i}) + (1-\alpha(\lambda,\zeta,\Omega_\mathrm{i}))\cdot T_m(\lambda,\zeta,\Omega_\mathrm{i})] \qquad (1)$$

where the hemispherical fluxes A, R, and T relate to the foliage absorption, black sky albedo, and canopy transmission measurements, respectively (given that no wood was present in the abstract canopy scenes). The background albedo, denoted here by α, was provided on the RAMI website. The overall deviation from energy conservation can be defined as the arithmetic average over a series of selected cases:

$$\Delta\mathrm{F}_m = \frac{1}{N_\mathrm{F}(m)} \sum_{\lambda=1}^{N_\lambda^m} \sum_{\zeta=1}^{N_\zeta^m} \sum_{i=1}^{N_{\Omega_\mathrm{i}}^m} \Delta\mathrm{F}_m(\lambda,\zeta,i)$$

where N_F is the total number of spectral λ, structural ζ, and illumination Ω_i conditions for which flux simulations were performed by model m.

Table 3. Deviation From Energy Conservation Expressed in Percent [%] for Test Cases With Lambertian Background Conditions in the Red and Near-Infrared (NIR) Spectral Bands. Indicated Are the Mean Difference ($\Delta\mathrm{F}$), the Maximum Difference ($\widehat{\Delta\mathrm{F}}$), the Standard Deviation ($\sigma_{\Delta\mathrm{F}}$), and the Fraction of Test Cases Performed (f_N)

Model Name	Lambertian RED				Lambertian NIR			
m	$\Delta\mathrm{F}$	$\widehat{\Delta\mathrm{F}}$	$\sigma_{\Delta\mathrm{F}}$	f_N	$\Delta\mathrm{F}$	$\widehat{\Delta\mathrm{F}}$	$\sigma_{\Delta\mathrm{F}}$	f_N
`FDM`	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.46
`DART`	0.00	0.00	0.00	0.84	0.07	0.38	0.09	1.00
`FLiES`	0.03	0.11	0.02	0.92	0.11	0.51	0.11	1.00
`parcinopy`	1.75	3.77	1.33	0.23	6.85	9.95	2.22	0.23
`raytran`	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
`RGM-2`	50.6	82.8	22.6	0.81	49.3	114.	34.1	0.88

[19] Table 3 provides the mean, the standard deviation, and the maximum value of $\Delta\mathrm{F}_m(\lambda,\zeta,\Omega_\mathrm{i})$ in percent [%] for simulations carried out in the red and NIR over test cases having Lambertian backgrounds. Energy conservation, in general, was well observed. One exception was the `RGM-2` model for which the absorption values were often close to unity such that a mean deviation of 50% from energy conservation was observed. Given the large differences between the absorption values of `RGM-2` and other models, it is appropriate to conjecture that the deviation from energy conservation is primarily due to operator errors rather than model deficiencies. Among the remaining models, `parcinopy` featured the largest energy deviation amounting on average to 1.75% in the red and 6.85% in the NIR. The only model to perform all of the red and NIR simulations and to adhere to energy conservation was `raytran`.

3.2. BRF Consistency

[20] This criteria relates to the fact that the sum of single-collided, single-uncollided, and multiple-collided BRF components must be equal to the total BRF, i.e., $\rho_{tot} = \rho_{co} + \rho_{uc} + \rho_{mlt}$. BRF consistency was found to hold true within 2×10^{-6} for all models and test cases with the exception of the heterogeneous canopies having anisotropic background conditions. For these cases, `librat` showed an average absolute deviation of 125×10^{-6} from BRF consistency. In addition, the multiple scattered BRF values provided by the `pbrt` model were sometimes negative in both the NIR and (more often) the red spectral domain. The largest observed negative value of `pbrt`'s multiple-scattered BRF component (–0.003821) occurred for simulations in the red spectral band along the PP above one of the heterogeneous canopies having anisotropic backgrounds (HET11) and a solar zenith angle of 50°. It is likely that this outcome is a consequence of deriving the multiple scattering components as the difference between the total BRF and the single (collided and uncollided) BRF components.

3.3. Spectral Consistency

[21] Spectral consistency relates to the fact that the ratio of the single-uncollided BRF component in two different wavelengths, $\rho_{uc}(\lambda_1)/\rho_{uc}(\lambda_2)$, must be equal to the

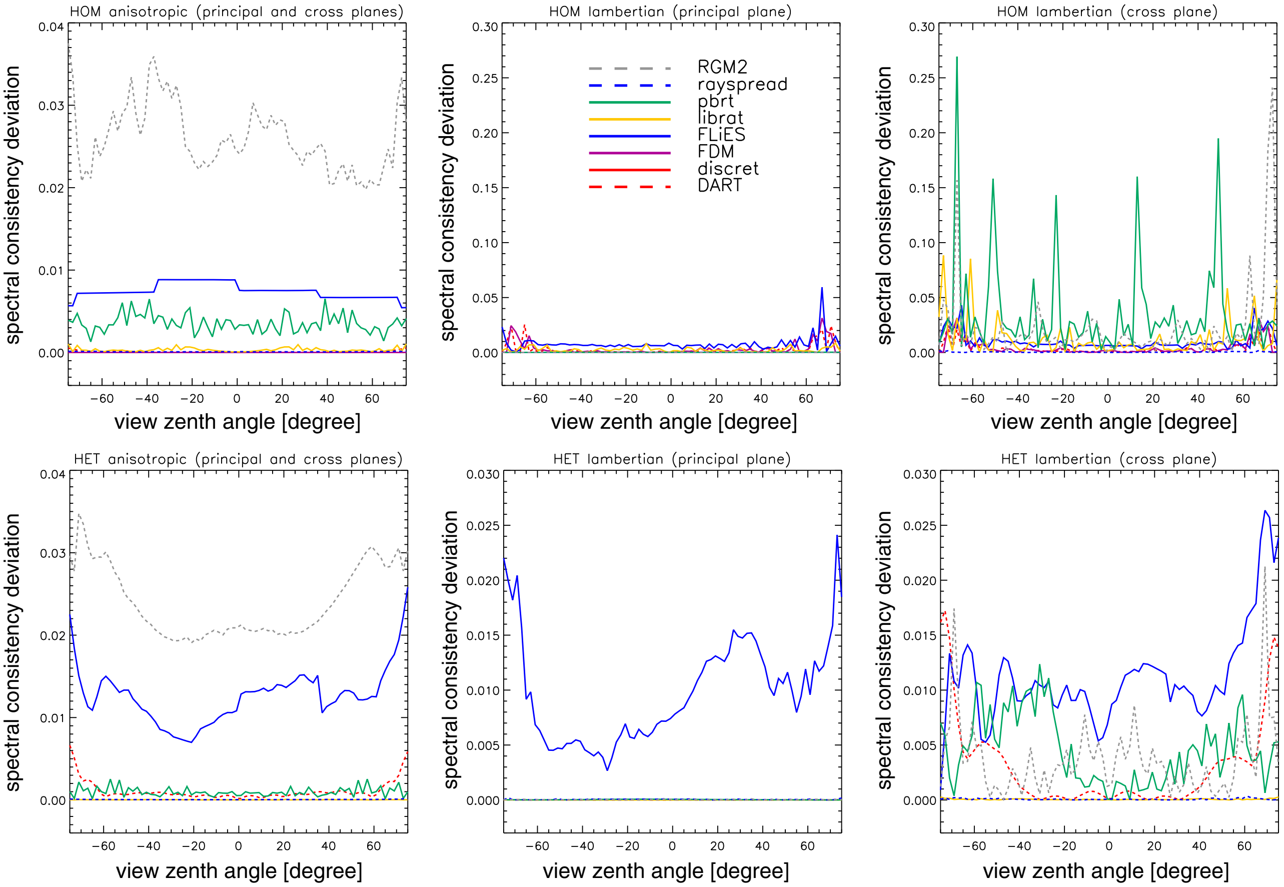

Figure 4. Graphs showing the mean absolute deviation from spectral consistency ($\langle|\Delta_S(\Omega_v)|\rangle$) derived from simulations of the single-uncollided BRF components in the red and NIR spectral bands for the homogeneous (top panels) and heterogeneous (bottom panels) "abstract canopy" test cases of RAMI-IV. The left graphs relate to simulations along both the principal and orthogonal planes for test cases with anisotropic background conditions. The middle and right panels relate to simulations along the principle and orthogonal planes, respectively, for test cases having Lambertian backgrounds.

ratio of the BRFs of the background of the canopy target $\rho_{bgd}(\lambda_1)/\rho_{bgd}(\lambda_2)$ in those two spectral regimes and for the same illumination (Ω_i) and viewing directions (Ω_v). For any model m the mean absolute deviation from spectral consistency over a variety of structural and illumination conditions ($N_S = N_\zeta^m \cdot N_{\Omega_0}^m$) can thus be defined as

$$|\Delta_S(m,\Omega_v)| = \frac{1}{N_S(m)} \sum_{\zeta=1}^{N_\zeta^m} \sum_{i=1}^{N_{\Omega_\mathrm{i}}^m} \left| \frac{\rho_{bgd}(\lambda_1,\zeta,\Omega_v,i)}{\rho_{bgd}(\lambda_2,\zeta,\Omega_v,i)} - \frac{\rho_{uc}^m(\lambda_1,\zeta,\Omega_v,i)}{\rho_{uc}^m(\lambda_2,\zeta,\Omega_v,i)} \right|$$

where for RAMI-IV test cases having Lambertian backgrounds, ρ_{bgd} is simply equal to the prescribed background albedo (α). Alternatively, for test cases with anisotropic background properties, $\rho_{bgd} = \rho_{bgd}^{\mathrm{RPV}}$, that is, the BRF of the background as generated with the RPV model.

[22] Figure 4 shows the mean absolute deviation from spectral consistency ($\langle|\Delta_S(\Omega_v)|\rangle$) as a function of view zenith angle for models having simulated the single-uncollided BRF component in the red and NIR. For test cases with anisotropic backgrounds (left panels), the pbrt model showed a bias of ~0.3% while that of FLiES was typically ~1% and that of RGM-2 around 2.5%. Interestingly, while the FLiES model remained at ~1% from spectral consistency for all Lambertian cases, the models pbrt and librat showed a noticeable increase in their deviations for simulations carried out along the OP over homogeneous test cases. Similarly, the models pbrt, RGM-2, and also DART displayed larger deviations along the OP (rather than the PP) for heterogeneous test cases. The largest deviations (up to 27% in the case of pbrt) occurred in the OP for homogeneous canopies having Lambertian backgrounds. In general, however, the observed biases were much smaller though. Since the deviations in Figure 4 occurred only for Monte Carlo models, they may have been caused by (1) insufficient numbers of rays, (2) differences in the seed values used by the random number generators in both spectral bands, and (3) differences in the ray numbers that were used when simulating BRFs in the red and NIR.

3.4. Model-to-Model Deviations

[23] This can be defined as the absolute normalized differences in the BRF (or flux) simulations between two models (c and m), when averaged over a variety of spectral

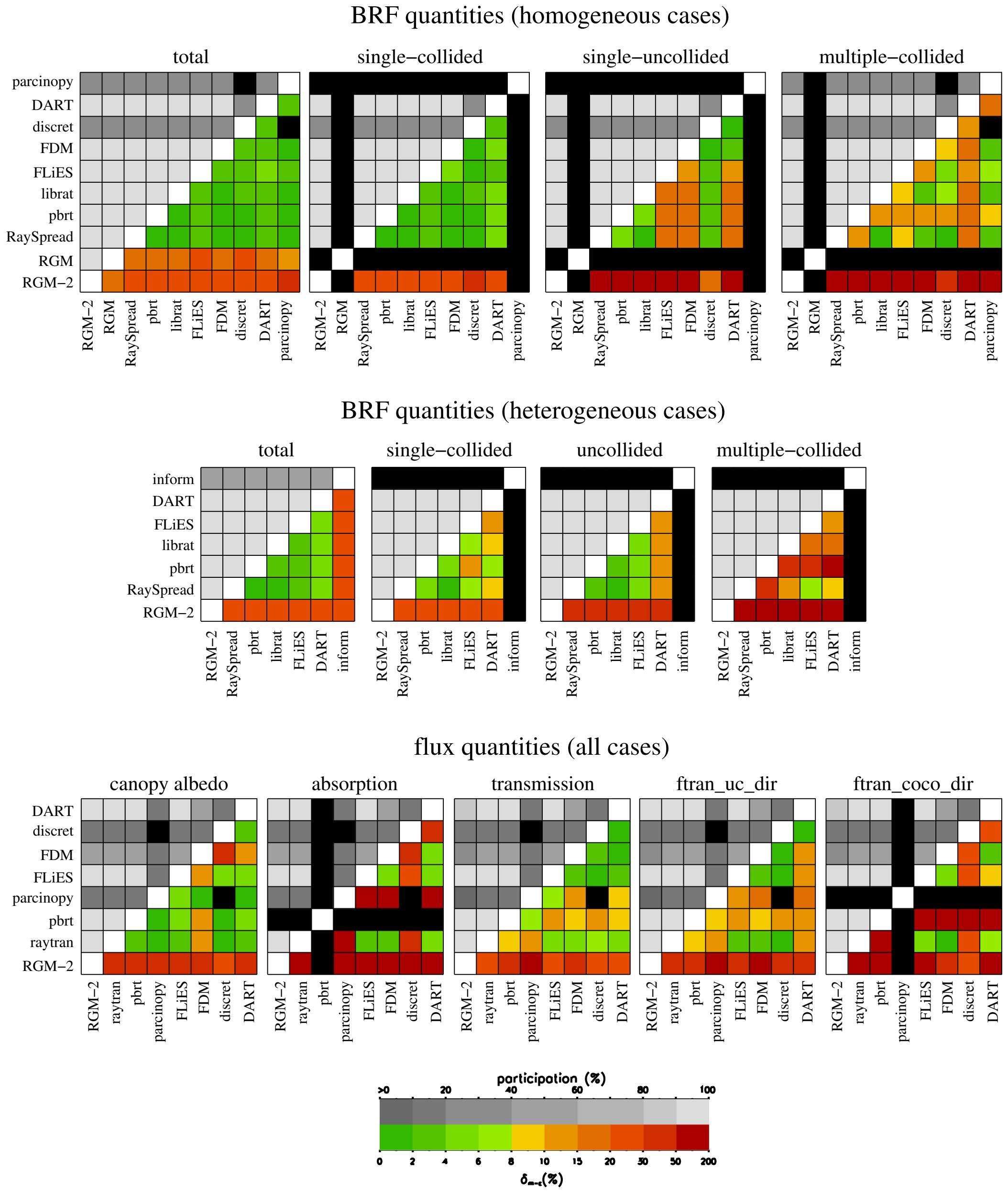

Figure 5. Model-to-model bias (lower right-hand side of graphs) and model participation (upper left-hand side of graphs) in percent for simulations of BRF and flux quantities. Shown are the total BRF, the single-collided by vegetation BRF, the uncollided by vegetation BRF, and the multiple-collided BRF over homogeneous (top row) and heterogeneous (middle row) canopy architectures. Also shown are the black sky albedo, the absorption by foliage, the total transmission, the uncollided transmission (ftran_uc_dir), and the collided-by-the-canopy-only component of the transmission (ftran_coco_dir) for all "abstract canopy" test cases (bottom row). Black boxes relate to model pairs without common simulations.

(λ), structural (ζ), viewing (Ω_v), and illumination (Ω_i) conditions:

$$\delta_{m\leftrightarrow c} = \frac{200}{N}\sum_{\lambda=1}^{N_\lambda}\sum_{\zeta=1}^{N_\zeta}\sum_{v=1}^{N_{\Omega_v}}\sum_{i=1}^{N_{\Omega_i}}\left|\frac{x_*^m(\lambda,\zeta,v,i)-x_*^c(\lambda,\zeta,v,i)}{x_*^m(\lambda,\zeta,v,i)+x_*^c(\lambda,\zeta,v,i)}\right|$$

where N is the number of simulations that have been performed by both models c and m. x_* relates to the (BRF or flux) target quantity, and $\delta_{m\leftrightarrow c}$ is expressed in percent.

[24] Figure 5 shows $\delta_{m\leftrightarrow c}$ (green to red colors in the lower right-hand side of a graph) together with the corresponding model participation (grey color scheme in upper left-hand side of graphs) in percent for simulations of BRF and flux quantities. Shown are model-to-model deviations for simulations of the total BRF (left column), the single-collided by vegetation BRF (second column), the uncollided by vegetation BRF (third column), and the multiple-collided BRF (right column) over homogeneous (top row) and heterogeneous (middle row) canopy architectures in both the red

and NIR spectral regimes. Black boxes relate to model pairs that do not have any simulations in common. Note that the results are normalized with respect to the number of BRF simulations that were carried out by a given pair of models.

[25] Mean absolute model-to-model BRF differences (top left panel) indicate a generally good agreement ($\delta_{m\leftrightarrow c} < 6\%$) between all of the participants apart from RGM and RGM-2. The latter two models stand somewhat apart for the total BRF simulations and, at least in the case of RGM-2, also for the various BRF components. This pattern is independent of the spectral band (not shown) or canopy scenario (although $\delta_{\mathtt{RGM-2}\leftrightarrow c}$ was less than 20% for the single-uncollided component over test cases with an anisotropic background as well as for the single-collided BRF component over the two layer canopy test cases). During RAMI-3, the RGM model delivered total BRF simulations for homogeneous canopies that did not deviate by more than a couple of percent from the majority of participating models. Given the radiosity nature of the RGM model, this could suggest that the observed differences in RAMI-IV are likely due to operator errors in the implementation of the new test cases. In fact, it turned out that the RGM-2 model simulations were based on architectural scenarios that differed from those prescribed on the RAMI website. For example, square leaves were used for the homogeneous test cases, a single spherical entity was used for the heterogeneous test cases instead of several thousand discs, and spatially reduced versions of the scenes were regularly used to save computing time. Since canopies are replicated indefinitely in these models, the latter simplification may have lead to artificial patterns of object arrangements (for the heterogeneous canopies) with subsequent biases in the simulated RT properties. This together with possible operator errors in the assigning of spectral canopy properties or the scaling of model-simulated radiative quantities may be the actual reasons for the observed biases of RGM-2.

[26] The models FLiES, FDM, and DART show deviations of up to 20% for the uncollided BRF component. At least for the DART model, this bias is very similar to that of the DART model participating in RAMI-3 (see Figure 6 in *Widlowski et al.* [2007b]). The large model-to-model differences of these three models may be affected by the very low uncollided BRF values of the HOM26/HOM36 two-layer canopy cases ($\rho_{uc} < 10^{-5}$ at large view zenith angles). In these scenarios, small differences in model simulations (for example, due to Monte Carlo noise) may strongly affect the value of $\delta_{m\leftrightarrow c}$. Model-to-model deviations generally increased for the multiple-collided BRF component and in particular so in the red spectral domain (not shown) due to the smaller BRF values there. Whereas the noise in the Monte Carlo simulations of pbrt (compare with the leftmost lower panel in Figure 2) could explain its increased $\delta_{m\leftrightarrow c}$ values, the FLiES and DART model simulations are not affected by noisy BRF signals yet they were found to differ from other models in the NIR (not shown). This was particularly so over homogeneous adjacent canopies and for canopies with an anisotropic background.

[27] Fewer models participated in the heterogeneous test cases of RAMI-IV. Again, the RGM-2 model is consistently different from other models here. The total BRF simulations of the inform model are also different from those of other RT models. DART simulations become increasingly different when going from the single-collided to the single-uncollided and finally to the multiple-collided BRF components. The bias of the uncollided BRF component arises from simulations pertaining to the constant slope and two-layer canopy cases in both the red and NIR. For the multiple-collided BRF simulations, it is again the simulations in the red spectral band (not shown) that lead to the largest deviations.

[28] The final row in Figure 5 shows the average model-to-model bias for the black sky albedo, the absorption by foliage, the total transmission, the uncollided transmission (ftran_uc_dir), and the collided-by-the-canopy-only component of the transmission (ftran_coco_dir). Whereas most models agree for simulations of the canopy albedo and transmission, this is no longer the case for the foliage absorption and the collided-by-the-canopy-only component of the transmitted flux. In particular, the discret and parcinopy models seem to differ in their canopy absorption estimates, whereas the pbrt and discret models stand relatively apart in their ftran_coco_dir simulations.

4. Proficiency Testing for RT Models

[29] The purpose of proficiency testing as described by ISO-13528 is "to demonstrate that the measurement results obtained by laboratories do not exhibit evidence of an unacceptable level of bias." The focus is thus not only on the "measurement"—which relates to the output of an instrument in response to external stimuli—but rather on the overall "method" that is used to obtain the measurement results. In general, the accuracy of the measurement method will depend on (1) the acquisition/preparation of the sample, (2) the appropriateness of the instrument's technology to deliver accurate results irrespective of the conditions under which the sample was acquired and subsequently analyzed, and (3) the choices/expertise of the operator carrying out the work (in a particular laboratory/outdoor environment).

[30] By analogy, the focus of RAMI is not only on the "simulation"—which relates to the output of a model in response to external inputs—but rather on the overall "method" that is used to generate the simulation results. In general, the accuracy of a simulation method depends on (1) the abstraction/representation of the target, (2) the appropriateness of the model's mathematical formulations to deliver accurate results irrespective of the nature of the target and the external forcings, and (3) the choices/expertise of the operator carrying out the work (in a particular computing language/environment). The purpose of using ISO-13528 in the context of RAMI is thus to demonstrate that the simulation results obtained by models do not exhibit evidence of an unacceptable level of bias.

4.1. Applying ISO-13528 to Canopy RT Models

[31] The following list describes how the various steps prescribed by ISO-13528 for inter-laboratory proficiency testing were implemented in the context of RAMI-IV:

[32] 1. *Ensure the homogeneity and stability of the samples that are to be analyzed by the participants.* These issues, while being relevant for interlaboratory proficiency tests, are essentially absent when it comes to RT model intercomparisons. This is so because the samples (called test cases in RAMI) are virtual and their characteristics are available in an exact, deterministic, and identical manner to all

RAMI-IV participants via the relevant webpages. It should be noted here that the RAMI coordinators have made every possible effort to provide detailed and accurate descriptions of the test cases and their associated measurement types. Should a model not be able to generate an exact copy of a RAMI test case, then this cannot be a limitation of the sample (i.e., the test cases provided on the RAMI website) but rather due to the model's internal RT formalism and/or the operator's choices when transferring the prescribed test case characteristics to the model. Although feasible, no efforts were undertaken in this phase of RAMI to randomize operator effects.

[33] 2. *Assign a reference value against which the bias of the participants can be determined.* Ideally, this assigned reference value (*X*) should come with a standard uncertainty (u_X) and be as close as possible to the true value of the quantity under study (here BRFs and hemispherically integrated radiative fluxes). In cases like RAMI where it is not possible to determine *X* and u_X prior to the launch of the model intercomparison exercise, ISO-13528 recommends to use consensus values derived either from the simulations of selected expert models or else from the participants of the proficiency test itself. These approaches are pursued here and described in greater detail in section 4.2.

[34] 3. *Specify a tolerance criteria allowing to determine whether deviations from the reference are significant.* Many evaluation metrics proposed by ISO-13528 include a measure of the bias levels that are still tolerable. For RAMI-IV, this "standard deviation of the proficiency assessment" ($\hat{\sigma}$) as it is called in ISO-13528 was prescribed in different ways depending on the radiative quantity of interest. For BRF quantities, it was expressed as a fixed fraction (*f*) of the reference (*X*):

$$\hat{\sigma}_{\rho*} = f \cdot X_{\rho*}$$

where the value of *f* was set to 0.03 and 0.05 in accordance with the 3–5% error margins obtained by vicarious calibration efforts of space borne remote sensing devices in the visible and NIR, e.g., *Thome* [2001], *Bruegge et al.* [2002], *Kneubühler et al.* [2002], *Thome et al.* [2008], and *Wang et al*. [2011].

[35] The proficiency standard deviation for canopy albedo (R) and foliage absorption (A) was defined using the maximum tolerable bias specified by the Global Climate Observing System (GCOS) in its satellite supplement to the implementation plan [*GCOS*, 2011]:

$$\begin{aligned} \hat{\sigma}_{\mathrm{R}} &= 0.05 \cdot X_{\mathrm{R}}/\sqrt{3} && \text{if} \quad 0.05 \cdot X_{\mathrm{R}} > 0.0025 \\ &= 0.0025/\sqrt{3} && \text{otherwise} \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_{\mathrm{A}} &= 0.10 \cdot X_{\mathrm{A}}/\sqrt{3} && \text{if} \quad 0.10 \cdot X_{\mathrm{A}} > 0.05 \\ &= 0.05/\sqrt{3} && \text{otherwise} \end{aligned}$$

where X_{R} and X_{A} are the reference values for the canopy albedo and foliage absorption, respectively, and the $\sqrt{3}$ factor arises from the (type B) rectangular distribution that was assumed in accordance with section 4.4 of *JCGM* [2008] for biases falling within the range tolerated by GCOS. Since canopy transmission is not an "essential climate variable" GCOS does not provide a corresponding tolerance criteria. However, if one assumes zero correlation between the various terms in the energy balance equation (equation (1)), then a proficiency standard deviation for T can be estimated on the basis of the GCOS accuracy criteria for A and R:

$$\hat{\sigma}_{\mathrm{T}} \approx \sqrt{\frac{\hat{\sigma}_{\mathrm{R}}^2 + \hat{\sigma}_{\mathrm{A}}^2}{(1 - R_{bgd})^2} + \frac{(1 - X_{\mathrm{R}} - X_{\mathrm{A}})^2}{(1 - R_{bgd})^4} \cdot \hat{\sigma}_{\mathrm{R}_{bgd}}^2}$$

For RAMI, the contribution from the background albedo is either zero (because $\mathrm{R}_{bgd} = \alpha$ is prescribed for Lambertian backgrounds) or else very small (numerical integration of RPV model). Hence, only the first term in the above equation was used to estimate $\hat{\sigma}_{\mathrm{T}}$. The resulting values of $\hat{\sigma}_{\mathrm{T}}$ were found to lie between 0.033 and 0.105 (mean value = 0.042) with little differences between the red and NIR spectral domain. To place this into perspective, the assigned reference values of the canopy transmission varied from 0.007 (0.205) to 0.885 (0.959) in the red (NIR) with an average value of 0.452 (0.676) over all abstract canopy cases in RAMI-IV.

[36] By setting the $\hat{\sigma}$ requirements in this manner, the proficiency assessment becomes equivalent to a "fitness for purpose" statement of the participating models, namely, whether they can simulate BRFs within the accuracies currently achieved by vicarious calibration efforts of satellite observations and whether they can match the GCOS accuracy criteria for canopy albedo and the fraction of absorbed radiation (FAPAR).

[37] 4. *Compare the uncertainty of the assigned reference value (u_X) with the tolerable deviation for the proficiency assessment ($\hat{\sigma}$).* ISO-13528 states that when $u_X \leq 0.3 \cdot \hat{\sigma}$, then the uncertainty of the assigned value is negligible and need not be included in interpretation efforts of the proficiency test [*ISO 13528*, 2005]. More specifically, if the above criteria are satisfied, then a simple *z*-score metric will suffice to assess model proficiency, while in cases where the standard uncertainty of the reference exceeds $0.3 \cdot \hat{\sigma}$, then z'-scores will have to be used as described in section 5.3. The actual computation of u_X is detailed in Appendix A and sections 4.2.3 and 4.2.4.

[38] If the standard deviation of the proficiency assessment ($\hat{\sigma}$) was assumed to be 3% (5%) of the assigned reference value, then the standard uncertainty of the assigned value was negligible in only 11.0% (26.7%) of the cases for the multiple-collided BRF component. At the same time, u_X was negligible for 87.8% (91.2%) of the single-uncollided BRFs, 99.2% (~100%) of the single-collided BRFs, and 71.2% (86.8%) of the total BRF simulations. Canopy structure affected the results for uncollided BRFs, while spectral bands affected those for multiple-collided and total BRFs. As such, u_X was negligible in only 51.4% (61.7%) of the uncollided BRF simulations pertaining to homogeneous two-layer canopies, whereas this fraction increased to almost 100% for heterogeneous canopies with anisotropic backgrounds if *f* = 0.03 (0.05). Similarly, the percentage of cases where u_X for the multiple-collided BRF component was compliant with the above ISO criteria changed from 1.5% (7.1%) for homogeneous canopies with anisotropic backgrounds to 34.7% (67.4%) for the homogeneous two layer test cases.

[39] For simulations of the canopy albedo, it turned out that the above $u_{X_{\mathrm{R}}} \leq 0.3 \cdot \hat{\sigma}_{\mathrm{R}}$ condition was satisfied on average for 59.7% of all cases. For canopy absorption, the compliance rate lay at 79.4%, while for transmission it was

78.3%. Results varied somewhat between test cases with the homogeneous two-layer canopy always delivering the best compliance (at almost 100%) while the heterogeneous two-layer cases had the least compliances, that is, 34.2% for R, 47.5% for A, and 51.7% for T. Overall, these results thus suggested that it is prudent to include the uncertainty of the assigned reference values in the data analysis step and to base any performance statistics on z'-scores rather than z scores.

[40] 5. *Evaluate the number of replicate simulations so that the repeatability standard deviation (σ_r) is negligible with respect to the tolerable deviation for the proficiency assessment ($\hat{\sigma}$).* Repeatability is defined here in analogy with its measurement counterpart in *JCGM* [2008], that is, as "the closeness of the agreement between the results of successive simulations of the same quantity carried out under identical conditions." ISO 13528 requires $\sigma_r/\sqrt{n} \leq 0.3\,\hat{\sigma}$ in order to eliminate the risk that repeatability variations cause the results of the proficiency test to be erratic (here n is the number of replicates). When it comes to physically based RT models, replicate simulations are rarely carried out. This is either due to the large computing times required to do so or because the majority of model classes identified by *Goel* [1988] delivers invariant simulation results for a given set of input values (assuming the computing environment is not modified between runs). Thus, $\sigma_r = 0$ for most RT models and the above ISO-13528 criteria is always satisfied. The sole exception to this comes from "Monte Carlo" (MC) ray-tracing models where convergence toward a solution is achieved at a rate proportional to the inverse square root of the number of rays that are used to stochastically sample the probability density function characterizing the scattering behavior of the canopy system under study [*Disney et al.*, 2000]. Conceptually, the repeatability of the simulations of a MC model will depend on (1) the number of rays that are used, (2) the seed number determining the locations of the incident rays above the target, (3) the architectural and spectral complexity of the system under study, and (4) the choice and degree of variance reduction techniques.

[41] The repeatability of MC ray-tracing models could not be determined using the "balanced uniform-level" experiments described in *ISO 5725-2* [1994] because a porting of MC models (acting as the "method" to be evaluated) to different computing environments (acting as "laboratories" here) was beyond the scope of RAMI-IV. Furthermore, since no two MC models in RAMI-IV made use of the same methodology for their ray propagation, weighing, and termination, it was decided to approximate the repeatability standard deviation by the within-model standard deviation s_W. For BRFs, where $\hat{\sigma}_{\rho*} = f \cdot X_{\rho*}$, the above ISO criteria can thus be rewritten as $s_W/(\sqrt{n} \cdot f \cdot X_{\rho*}) \leq 0.3$. This allows evaluation of the compliance of MC models using two different approaches depending on whether the models generated smooth or noisy BRF datasets (or alternatively whether they made use of variance reduction techniques or not). For the latter set of MC models, s_W^2 was determined as the arithmetic mean of three intermediate precision variances s^2 (as defined in equation 10 of *ISO 5725-3* [1994]). More specifically, the s^2 values were computed from $n = 26$ BRF simulations along the orthogonal plane for test cases where the BRFs were essentially invariant over a range of view zenith angles. An example of such a BRF dataset can be found in the top left panel of Figure 2 for view zenith angles between ±25°. Using this approach on the models parcinopy, pbrt, RGM, and RGM-2, it was found that $s_W/(\sqrt{n} \cdot 0.03X_{\rho*})$ was less than 0.3 except for pbrt simulations of the multiple-collided BRF component in the red spectral band (compare with the green data in the lower rightmost graph of Figure 2).

[42] MC models that delivered predominantly smooth BRF data sets could not be evaluated in this manner. Instead, these models were asked to provide BRF simulations for one of the most heterogeneous two-layer canopy test case (HET27). More specifically, for a solar zenith angle at 20°, the models were to generate BRFs using different starting seeds for the ray-tracing process and also several levels of ray numbers. Figure 6 shows the mean value of $s_W/(0.03X_{\rho*}\sqrt{n})$—computed from all BRFs in the PP—plotted on a log-log scale against the number of rays that were used in the simulations. The colored discs indicate the number of rays that a MC model used when generating the BRFs (FLiES, pbrt, rayspread) or fluxes (raytran) for the RAMI-IV exercise. The area of validity of the ISO criteria is shaded in grey, and different panels in Figure 6 relate to different BRF components (columns) and spectral domains (rows). All models exhibit improvements in the standard deviation that are in line with the theoretical $1/\sqrt{raynumber}$ relationship (i.e., the slopes in Figure 6 are close to –0.5). Only the simulations of the FLiES model and the pbrt model for the multiple scattered BRF component in the red spectral domain do not comply with the ISO-13528 criteria (discs are above grey area). Any "action" or "warning" flags that may arise for these models and simulation conditions later on should thus be interpreted carefully since these may actually be due to insufficient numbers of incident rays rather than deficiencies in the models themselves.

[43] 6. *Compute performance statistics that document the proficiency of the participating models.* ISO-13528 provides eight different statistical metrics to analyze the results of participants. Out of these performance measures, three metrics could not be used because of the significant uncertainty associated with some of the assigned reference values (see item 4 above). Others could not be used because the participating models in RAMI-IV did not provide standard uncertainties of their simulations. As a result, section 5 will first report on the bias statistics between model and reference simulations. This will be followed by E_n numbers and z'-scores.

4.2. Assigning Reference Values

[44] One of the main challenges in the verification of canopy RT models is the general lack of comprehensive reference datasets. While it is conceptually impossible to measure the true value of a target quantity in the field or laboratory [*JCGM*, 2008], this is no longer the case when dealing with perfectly controlled virtual environments. In such designer-based systems it becomes possible, under certain well defined conditions, to determine the true value of the target quantity with the help of exact analytical solutions. In these select cases, the performance of canopy RT models can thus be evaluated with respect to an absolute truth. Using such an approach in RAMI-3 allowed to identify a series of six "credible" 3-D Monte Carlo models [*Widlowski et al.*, 2007b]. More specifically, these models were able to match a series of analytical solutions to within the precision levels required by RAMI. In addition, they showed an over-

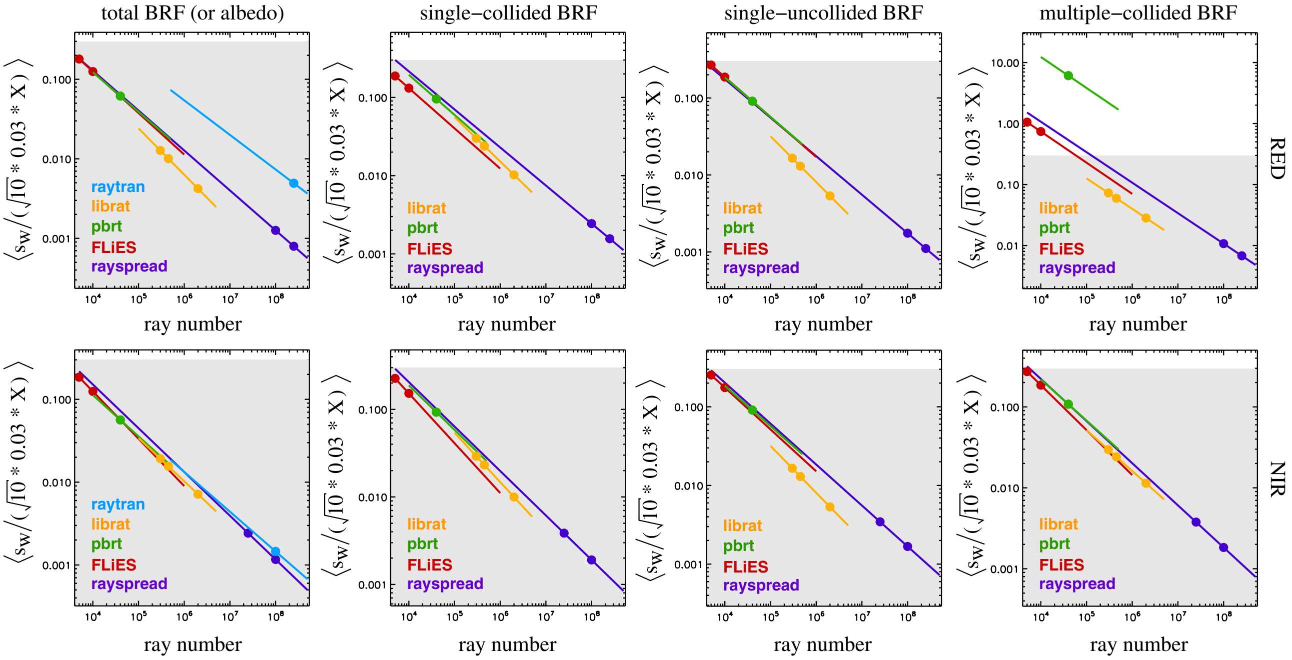

Figure 6. Log-log plots of normalized standard deviations ($s_W/\sqrt{n}\cdot\hat{\sigma}$)—derived from BRF (or DHR) simulations of the HET27 test case having a solar zenith angle of 20°—as a function of the number of rays that were used by the MC models. The proficiency standard deviation was expressed as 3% of the assigned reference BRF, i.e., $\hat{\sigma}\ =\ 0.03X$ and $n\ =\ 10$ is the number of BRF replicates (different seed values) at a given ray number. The discs indicate the number of rays used by a given model (color) when performing RAMI-IV experiments. The grey area indicates compliance with the ISO-13528 criteria. Only the raytran data relate to surface albedo (DHR) simulations.

all agreement of ∼1% across all of their BRF simulation results. Four of these "credible" RAMI-3 MC models contributed to RAMI-IV, namely, DART, librat, raytran, and rayspread. Of these, the DART model had been substantially altered since RAMI-3. DART now makes use of a flux tracking method (rather than MC ray-tracing) such that it cannot be considered identical to the model analyzed in *Widlowski et al.* [2007b].

[45] In the context of RAMI-IV, the true values of the radiative quantities are not known *a priori*. For proficiency testing, one thus has to make use of section 5.5 (consensus value from expert laboratories) and/or section 5.6 (consensus value from participants) of ISO-13528 to assign reference values. As indicated previously, the expert models that will be used in RAMI-IV are the "credible" MC models librat and rayspread for BRF simulations and the raytran MC model for the flux simulations. The following sections will describe in more detail how the assigned reference values (denoted X) were derived from the available model simulations (denoted ρ for BRFs).

4.2.1. Single-Collided and Uncollided BRFs

[46] The first row of Figure 7 displays histograms of the relative biases between librat and rayspread simulations carried out over all of the RAMI-IV actual canopy scenarios. In the case of the single-collided BRF component (leftmost panel) and the single-uncollided by vegetation BRF component (middle panel), the relative differences between both MC models are narrowly distributed with a mean bias of 0.03% ±0.73% and 0.21% ±3.23%, respectively. The somewhat larger spread of the uncollided BRFs is due to the very low values obtained for some of the homogeneous two layer canopies (where $\rho_{uc}\approx 10^{-5}$). In this case, small differences between the model simulations (carried out with a precision of 10^{-6}) will lead to inflated relative differences. The middle row of Figure 7 provides scatter plots of the same set of librat and rayspread simulations. Both the signal to noise ratio (SNR = 334.6 and 572.0) and the linear regression coefficients (R^2= 0.99998 and 0.99999) confirm the very good agreement between the two MC models (for single-collided and single-uncollided BRF simulations, respectively).

[47] A good agreement between two data sets is, however, not yet proof of their accuracy. The latter requires a direct comparison of at least one of the data sets with a reference having a known bias/precision. More specifically, the uncertainty of the reference should be smaller/better than that of the candidate dataset. In a second step, the thus "calibrated" candidate dataset can then act as a transfer standard to characterize the quality of the remaining one. This process, which is known as a "metrological traceability chain" [*JCGM*, 2012], can also be applied to MC ray-tracing models if they utilize deterministic representations of their canopy targets and no undue shortcuts in the RT simulations. To this end, the two leftmost panels in the third row of Figure 7 document the verisimilitude between rayspread simulations from RAMI-3 and the corresponding exact (analytical) solutions for ρ_{co} (first column) and ρ_{uc} (second column). For these two BRF components, it is

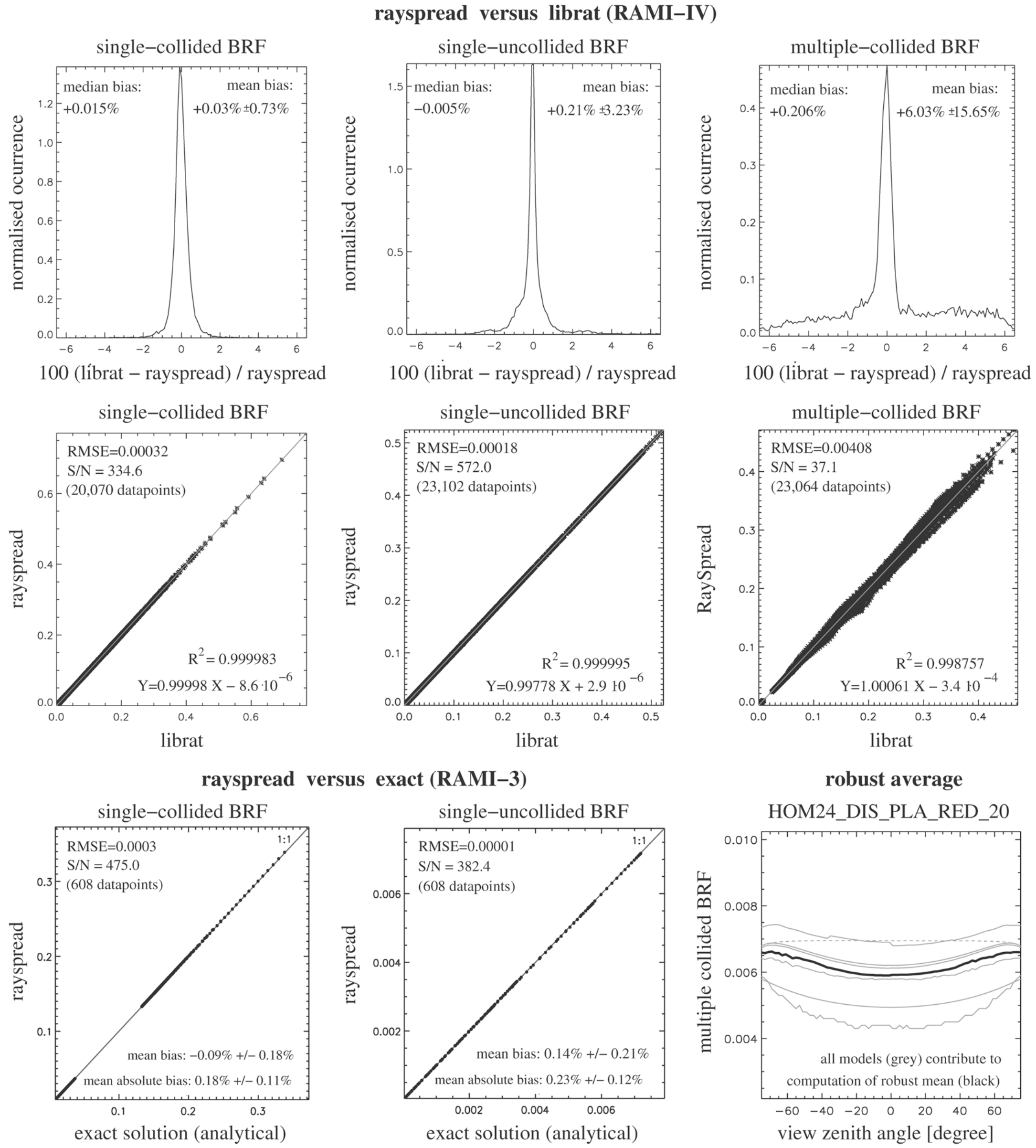

Figure 7. Results in columns relate (from left to right) to simulations of the single-collided, the single-uncollided, and the multiple-collided BRF. The top row shows histograms of the relative differences occurring between `rayspread` and `librat` simulations over the complete set of RAMI-IV abstract canopy test cases. The middle row depicts scatter plots of the same simulation results. The bottom row provides an example of a robust average for multiple-collided BRF data (rightmost panel) as well as scatter plots between `rayspread` simulated—single-collided (first column) and single-uncollided (second column)—BRFs and the corresponding exact (analytical) solutions for a set of homogeneous turbid medium canopies with uniform LNDs taken from RAMI-3.

possible to determine the analytical solutions over homogeneous turbid medium canopies with uniform leaf normal distributions (LND) and Lambertian scattering properties. This thus allows to determine the model bias exactly. For the `rayspread` simulations of ρ_{co}, the bias turned out to be –0.09% ±0.18%, while for ρ_{uc}, it was 0.14% ±0.21%. Combining these findings with the results obtained from the RAMI-IV comparison between `rayspread` and `librat` yields a mean bias value (between `librat` and the exact, *i.e.*, analytical, solution) of –0.06% ±0.75% for ρ_{co} and 0.35±3.24% for ρ_{uc}.

[48] Given the proficiency testing criteria of 3 to 5% for BRFs, the above results document that both the `librat` and `rayspread` MC models are sufficiently close to each other and to the exact (analytical) reference solutions to serve as the baseline for assigning the reference values for the single-collided (X_{co}) and uncollided (X_{uc}) BRF components in RAMI-IV. The determination of the final reference value is carried out using the robust analysis algorithm proposed in annex C of ISO-132528 (and outlined in Appendix A here). More specifically, when based on two models, the robust mean is simply the arithmetic mean of the two input

data. However, because the ρ_{co} simulations of librat for homogeneous canopies with anisotropic backgrounds in the red spectral band were resubmitted with a relatively large noise level, it was decided to base the reference solution for these cases on the simulation results of the rayspread model only.

4.2.2. Multiple Scattered BRFs

[49] The rightmost panels in the first row of Figure 7 displays a histogram of relative differences in the multiple-collided BRF simulations of librat and rayspread for the abstract canopy test cases of RAMI-IV. The mean bias between the two MC models is 6.03% ± 15.65% (using 23,064 data points) with some values in the red spectral band (where ρ_{mlt} is often rather low) reaching 50% bias or more. The larger spread in the ρ_{mlt} simulations (compared to the ρ_{uc} and ρ_{co} BRF components) is also visible from the scatterplot in the rightmost panel of the middle row of Figure 7. Here the SNR=37.1 which is only about a tenth of what it was for ρ_{co} and ρ_{uc}. The reasons for these differences are not absolutely clear at this stage although a preliminary analysis points toward an erroneous post-processing of some the librat ρ_{mlt} simulations. While the observed mean amplitude of the librat simulated BRF signal appears to be similar to that of other RT models in RAMI-IV, the angular shape of its ρ_{mlt} component is often different in particular along the principal plane.

[50] Due to the lack of exact analytical solutions for ρ_{mlt} and since the typical bias between the librat and rayspread models exceeded the prescribed criterion for the proficiency test (i.e., $\hat{\sigma}_{\rho_{mlt}} = 0.03X_{mlt}$ or $0.05X_{mlt}$), it was decided to adopt the approach of section 5.6 in ISO-13528 (consensus values from participants) to assign reference values for the multiple-collided BRF component (X_{mlt}). Data sets were excluded from the robust analysis—described in Appendix A—if their noise level (or angular pattern) perturbed the smoothness of the robust mean along the principal or orthogonal planes. This concerned the ρ_{mlt} simulations of (1) pbrt in the red spectral band (all cases) as well as for the constant slope test cases and for the homogeneous test cases in the NIR, (2) parcinopy for the homogeneous two-layer canopies having a planophile LND below an erectophile foliage layer, and (3) RGM-2 for the homogeneous two layer test cases HOM27 and HOM28 in the red spectral band for $\theta_0 = 50°$. Despite these exceptions, the number of RT model simulations that were used to compute the robust average stayed between 4 and 7. An example of the outcome of such a robust analysis is shown in the rightmost panel in the bottom row of Figure 7.

4.2.3. Total BRFs

[51] The robust analysis was not applied to the total BRF simulations that were provided by the RAMI-IV participants. Instead, it was decided to define the assigned reference value for total BRFs (X_{tot}) as the sum of the assigned reference values of the three BRF components, i.e., $X_{tot} = X_{uc} + X_{co} + X_{mlt}$. Similarly, the variance of X_{tot} was computed as the sum of the variances of the relevant reference values: $u^2_{X_{tot}} = u^2_{X_{uc}} + u^2_{X_{co}} + u^2_{X_{mlt}}$. This approach ensures consistency.

[52] It should be noted here that ISO-13528 prohibits the evaluation of models/laboratories if their data were used in the generation of the assigned reference values. To avoid such correlation issues within RAMI, every participating model had its own reference value computed. This was done by excluding the model in question from the list of models contributing to the robust mean. The assigned reference value (X_*) thus becomes model-specific (X^m_*) in this study. In the case of ρ_{co} and ρ_{uc}, for example, this meant that the assigned reference values for rayspread and librat differed from that computed for the remainder of models, whereas for ρ_{mlt}, the assigned reference value was different for every model that contributed to the robust analysis. The differences between the assigned reference BRFs were on average less than 0.5% for ρ_{co}, ρ_{uc}, and ρ_{tot} but could reach up to 3% for ρ_{mlt}.

[53] The robust analysis approach proposed by ISO-13528 delivers also an estimate of the standard uncertainty of the reference (u_X) (see equation A5). For cases where the reference had to be based on a single "credible" model—e.g., when using the librat model as reference for ρ_{co} and ρ_{uc} simulations of rayspread (and vice versa)—then the value of u_X had to be derived by other means. More specifically, in such cases, the standard uncertainty was estimated as $u_X = s_W/\sqrt{n}$ using the BRF data contributing to Figure 6. Here s_W was the average of the standard deviations obtained from $n = 10$ different sets of BRF simulations for each viewing condition along the PP over the HET27 heterogeneous two layer canopy test case. This process was carried out using the value of s_W corresponding to the smallest number of incident rays that were used by the librat and rayspread models to perform the RAMI-IV simulations. The resulting standard uncertainty values were $u^{\texttt{librat}}_{\rho_{uc}} = 1.39\cdot10^{-5}$ $(1.71\cdot10^{-5})$ and $u^{\texttt{rayspread}}_{\rho_{uc}} = 1.71\cdot10^{-6}$ $(4.01\cdot10^{-6})$ in the red (NIR) and $u^{\texttt{librat}}_{\rho_{co}} = 8.63 \cdot 10^{-6}$ $(7.22 \cdot 10^{-5})$ and $u^{\texttt{rayspread}}_{\rho_{co}} = 7.61 \cdot 10^{-7}$ $(1.04 \cdot 10^{-5})$ in the red (NIR).

4.2.4. Hemispherical Fluxes

[54] The raytran model was the only "credible" MC model to deliver hemispherically integrated fluxes in RAMI-IV. During RAMI-3, it had been shown to match analytical solutions of both the foliage absorption and canopy albedo to within the numerical precision required by RAMI (10^{-6}). In section 3, raytran was furthermore shown to conserve energy (for test cases with Lambertian backgrounds). As such, it was considered appropriate to make use of the simulations of raytran to assign reference values for hemispherically integrated fluxes in RAMI-IV. In this case, the standard uncertainty of the reference was again estimated as $u_X = s_W/\sqrt{n}$ where s_W was the standard deviation of $n = 10$ flux simulations carried out with different starting seeds and for ray numbers typically used by raytran for its RAMI-IV flux simulations (compare with the albedo data of raytran in Figure 6). As such, $u_R = 5.18\cdot10^{-6}$ in the red and $1.37\cdot10^{-5}$ in the NIR. At the same time, and in order not to rely solely on the simulations of a single RT model, it was decided to generate a second set of reference values based on a robust analysis of the hemispherical fluxes generated by all of the RAMI-IV participants.

5. Results for "Abstract Canopy" Cases

5.1. RT Model Bias

[55] A detailed analysis of the patterns emerging when total BRF differences, i.e., $\rho_{tot} - X_{tot}$, were plotted as a function of the reference value (not shown) allowed to draw the following conclusions:

[56] 1. The bias of the models `librat`, `rayspread`, and `pbrt` can be considered independent of the reference value in both the red and also the NIR spectral domain. Some larger deviations were noted for $\rho_{tot}^{\texttt{pbrt}}$ over heterogeneous test cases with a constant inclination of crowns. These deviations are likely to be due to operator glitches since the bias was directionally selective (as can be seen from the second panel of the bottom row in Figure 2).

[57] 2. The bias of the models `DART`, `discret`, `FDM`, `FLiES`, and `parcinopy` are likely to be independent of the reference value in at least one (typically the red) spectral domain. Only the range of the biases of $\rho_{tot}^{\texttt{FLiES}}$ in the NIR seemed to increase consistently with the reference BRF.

[58] 3. The bias of the models `inform`, `RGM`, and `RGM-2` is often rather large and does not exhibit any clear pattern with respect to the assigned reference values. Typically, the bias is smaller in the red band and also evenly distributed around the zero line in that spectral regime. At the same time, all three models have a tendency to systematically underestimate the reference value in the NIR. This spectral pattern is likely to be caused by an underestimation of ρ_{mlt} in the NIR.

[59] It should be noted that this qualitative analysis of bias patterns does not account for uncertainties associated with the model simulations and/or reference values. Neither does it relate the observed biases to the prescribed standard deviation of the proficiency test. The following sections will deal with these aspects in order to determine whether the observed differences between model and reference values are actually relevant.

5.2. $\mathbf{E_n}$ Numbers

[60] The E_n performance statistics is suggested in section 7.5 of ISO-13528 to evaluate the reliability of the expanded uncertainty that individual laboratories claim to have. In the context of RAMI, it is defined as

$$E_n(m;\lambda,\zeta,\Omega_v,\Omega_{\rm i}) = \frac{x_*^m(\lambda,\zeta,\Omega_v,\Omega_{\rm i}) - X^*(\lambda,\zeta,\Omega_v,\Omega_{\rm i})}{\sqrt{U^2_{x_*^m}(\lambda,\zeta,\Omega_v,\Omega_{\rm i}) + U^2_{X_*}(\lambda,\zeta,\Omega_v,\Omega_{\rm i})}} \tag{2}$$

where x_*^m is the radiative quantity of interest (e.g., a total BRF, a flux quantity, or one of their components) simulated by model m for a given spectral (λ), structural (ζ), viewing (Ω_v), and illumination ($\Omega_{\rm i}$) related condition. X_* is the assigned reference value under these conditions (which is model-specific here, i.e., X_*^m), while the expanded uncertainty of the reference $U_{X_*} = k \cdot u_{X_*}$ with $k = 2$ as coverage factor. The standard uncertainty of the reference u_{X_*} was defined either as in equation (A5) or, for cases where the assigned reference value originates from a single model, as specified in sections 4.2.3 and 4.2.4. A value of $|E_n| < 1$ provides objective evidence that the estimates of expanded uncertainty agree with the observed differences between x_*^m and X_*. Alternatively, one can say that if $|E_n|$ is less than unity, then the expanded uncertainties of both the model and reference data are sufficient to explain the differences observed between them. Two different usages of E_n statistics will now be presented.

5.2.1. Maximum Tolerable Model Uncertainties

[61] A first usage of E_n statistics requires the assumption that operator errors are absent and that the expanded uncertainty of the RT model is equal to the maximum tolerable uncertainty level for a given target quantity. In other words, the expanded uncertainty for the model $U_{x_*^m} = k \cdot u_{x_*^m}$ where $k = 2$ and the standard uncertainty of the model $u_{x_*^m}$ is equal to $\hat{\sigma}_*$ defined in item 3 of section 4.1. Under these conditions, an $|E_n|$ value greater than unity thus implies that the model simulated BRF or flux value deviates by more than the typical uncertainties associated with satellite observations or by more than the accuracy criteria specified by GCOS from the assigned BRF or flux reference value, respectively. After applying equation (2) to all (~10,000) BRF or (76) hemispherical flux simulations that a model could generate in the red or NIR spectral domain, the resulting $|E_n|$ values were arranged in order of increasing magnitude. From this sequence, the $|E_n|_{\rm P}$ value at a specific percentile P can then be selected to report on the model's performance. For example, if P=50 (100), then $|E_n|_{50}$ ($|E_n|_{100}$) would report on the median (maximum) $|E_n|$ value of all BRF/flux quantities generated by a given model.

[62] The top row in Figure 8 displays graphs of the 95th percentile of $|E_n|$ in the red spectral band plotted against the same quantity ($|E_n|_{95}$) in the NIR for different BRF components (top row). Every model (colour) is shown by two points, namely, when the standard uncertainty of the model ($u^m_{\rho*} = f \cdot X_{\rho*}$) was set to 3% (triangle) or 5% (pentagon) of the reference value $X_{\rho*}$. When $|E_n|_{95} < 1$, the uncertainties associated with both the model and reference values are larger than their respective differences such that at least 95% of the model-simulated data can be considered equivalent to the reference data. This condition applies to the single-collided BRF simulations of `discret`, `FDM`, `librat`, and `rayspread` in both the red and NIR irrespective of whether one assumes 3 or 5% errors in the model. For simulations of ρ_{uc}, only the "credible" MC models `librat` and `rayspread` are indiscernible from the reference at both 3 and 5%. At the same time, however, the `discret` model reaches $|E_n|_{95} < 1$ if $u_{\rho_{uc}} = 0.05 \cdot X_{\rho_{uc}}$. Overall, little bias exists between model performances in the red and NIR spectral domain for ρ_{co}. Once exception perhaps is the `librat` model which deviates somewhat from the one-to-one line for ρ_{co} simulations. This is due to the large noise level in its (resubmitted) ρ_{co} simulations for homogeneous canopies with anisotropic background properties (that were therefore excluded from the reference solution). For simulations of ρ_{mlt}, the models `discret`, `rayspread`, and `parcinopy` are indiscernible from the assigned reference solutions at both $f = 0.03$ and 0.05. A series of models have $|E_n|_{95} < 1$ in the NIR but not in the red. One of these is the `librat` model, where it had been noted that the shape of some of its ρ_{mlt} simulations along the PP showed unusual dips. The large $|E_n|_{95}$ values for models `pbrt` (and to a lesser extent also `FLiES`) in the red spectral domain are likely to be due to the large repeatability uncertainty of their ρ_{mlt} simulations (see top right panel of Figure 6). This argument, however, does not apply to `DART` which exhibits a 5 times larger $|E_n|_{95}$ value for ρ_{mlt} in the red than in the NIR.

[63] The choice of percentile level impacts the location of a model in the above type of graphs. Hence, it is of interest to document in what manner a model may migrate (or not) from the $|E_n|_{\rm P} \leq 1$ regime to the $|E_n|_{\rm P} > 1$ regime as P → 100. This can be envisaged from the panels in the middle row of Figure 8 which show the percentage of model simulations in the red spectral domain for which $|E_n| > 1$ (plotted

Table 4. Absolute ($U_{95}^{\rho_{tot}}$) and Relative ($\tilde{f}_{95}^{\rho_{tot}}$) Values of the Expanded Uncertainty Associated With a Given RT Model Such That 95% of All Its Total BRF Simulations (ρ_{tot}) Become Indistinguishable From the Reference Solution (Such That on Average $|E_n| = 0.5$)[a]

Model	λ = red			λ = NIR		
Name	$U_{95}^{\rho_{tot}}$	$\tilde{f}_{95}^{\rho_{tot}}$	f$_N$	$U_{95}^{\rho_{tot}}$	$\tilde{f}_{95}^{\rho_{tot}}$	f$_N$
DART	0.01349	28.9%	100.0%	0.06489	18.4%	100.0%
discret	0.01617	10.3%	15.8%	0.03601	5.80%	15.8%
FDM	0.00571	8.13%	47.4%	0.06005	9.21%	47.4%
FLiES	0.01263	15.7%	100.0%	0.06759	13.0%	100.0%
inform	0.08603	57.9%	31.6%	0.39287	na	31.6%
librat	0.00091	na	99.9%	0.01848	na	99.8%
parcinopy	0.00161	7.6%	15.8%	0.03203	6.06%	15.8%
pbrt	0.00567	na	100.0%	0.02601	na	100.0%
rayspread	0.00043	na	84.2%[b]	0.00507	na	99.8%
RGM	0.01700	26.3%	46.7%	0.55498	87.4%	46.7%
RGM-2	0.09831	na	99.3%	0.45893	na	99.3%

[a] "Not applicable" (na) refers to cases where the model bias was not dependent on the value of the reference. The fraction of cases a model simulated is also indicated (f$_N$). Reference BRF values in the red span the range 0.0 to 0.5, while those in the NIR cover 0.2 to ~1.1.

[b] rayspread simulated all of the prescribed BRFs, but since some of the librat simulations, which act as reference for rayspread, had to be excluded due to their increased MC noise levels, the number of cases reported here for rayspread is less than 100%.

against the same statistics in the NIR). Results are shown for both $u_{\rho_*}^m = 0.03 \cdot X_{\rho_*}$ and $u_{\rho_*}^m = 0.05 \cdot X_{\rho_*}$. The vertical (horizontal) black line segment delineates the 95th percentile in the NIR (red) such that models which fall to the left of (below) this line will have $|E_n|_{95} < 1$ in the NIR (red) in the corresponding top row graphs. Thus, by lowering the tolerable percentage of BRF cases with $|E_n| > 1$ to 0.1 one will observe, for example, that only one model per BRF component (i.e., rayspread for ρ_{co} and ρ_{uc} and no model for ρ_{mlt}) will have $|E_n|_{99.9} \leq 1$ in both the red and NIR if $f = 0.03$. If the model uncertainty is relaxed to $f = 0.05$, then also the discret model (for ρ_{co}) and the parcinopy model (for ρ_{mlt}) will have $|E_n|_{99.9} \leq 1$ in both the red and NIR. For ρ_{mlt} simulations, the models discret, rayspread, and parcinopy achieve $|E_n|_{99.9} \leq 1$ at $f = 0.03$ but only in the NIR. Similarly, for ρ_{co}, the model librat (parcinopy) obtains this result at $f = 0.03$ ($f = 0.05$). On the other end of the performance scale lies RGM-2, which never achieves $|E_n| \leq 1$ for ρ_{mlt} simulations in the NIR.

[64] Given the small number of hemispherical flux simulations available (at best 76 values per spectral band), the bottom row of Figure 8 displays again the 95th percentile of $|E_n|$ in the red versus that in the NIR. More specifically, results are shown for different hemispherical fluxes (columns) and different methods adapted when assigning the reference value (symbol shapes). Note that when the robust analysis approach (see Appendix A) was used to assign the reference values, the simulations of the RGM-2 model were excluded due to its large deviations from energy conservation (see section 3). Canopy absorption estimates for the model pbrt were derived assuming a perfect closure of the energy balance equation. One will note that several models are capable of matching the GCOS accuracy criteria in both the red and NIR spectral domains (at least for 95% of their simulations), while others can only do so in one of the spectral regimes. At the same time, it is apparent that the choice of reference solution (i.e., raytran simulations versus a robust mean approach) has only a limited impact on the model results in these graphs. Last but not least, when choosing raytran simulations as reference, no major biases were noted between model simulations in the red and NIR.

5.2.2. Actual Expanded Model Uncertainties

[65] An alternative usage of E_n statistics is to re-arrange equation (2) such that for any given simulation result, it becomes possible to compute the value of the expanded uncertainty of the model ($U_{x_*^m}$) that makes the bias ($x_*^m - X_*$) acceptable:

$$
\begin{array}{ll}
U_{x_*^m} = \sqrt{\dfrac{(x_*^m - X_*)^2}{E_n^2} - U_{X_*}^2} & \text{if} \quad U_{X_*}^2 < \dfrac{(x_*^m - X_*)^2}{E_n^2} \\
U_{x_*^m} = 0.0 & \text{otherwise}
\end{array}
$$

where the $(\lambda, \zeta, \Omega_v, \Omega_i)$ notation has been dropped for ease of reading. By setting $E_n = 0.5$ in the above equations (which is equivalent to assuming that $|E_n|$ is uniformly distributed in the range 0 to 1), one can compute $U_{x_*^m}$ (or $U_{x_*^m}/X_* = \tilde{f}_{x_*^m}$) for every simulated BRF or flux in RAMI-IV. The choice between $\tilde{f}_{x_*^m}$ or $U_{x_*^m}$ depends on the pattern exhibited by the biases reported in section 5.1. For models where the BRF bias was increasing with the value of the reference BRF, it is appropriate to report $\tilde{f}_{x_*^m}$, whereas for those models where the bias was more or less a constant, the reporting of $U_{x_*^m}$ is more appropriate.

[66] From among the 10,000 or so $U_{x_{\rho_{tot}}^m}$ (or $\tilde{f}_{x_{\rho_{tot}}^m}$) values that were computed from total BRF simulations in the red or NIR, Table 4 reports only the expanded uncertainty which allows explaining 95% of the observed biases, i.e., $U_{95}^{\rho_{tot}}$ (or $\tilde{f}_{95}^{\rho_{tot}}$). Ignoring librat and rayspread, the model with the lowest expanded uncertainty in the red spectral domain was parcinopy ($U_{95}^{\rho_{tot}} = 0.00161$ with a participation rate of 15.8%) followed by pbrt ($U_{95}^{\rho_{tot}} = 0.00567$ with a participation rate of 100%). In the NIR, it was pbrt ($U_{95}^{\rho_{tot}} = 0.02601$ with a participation rate of 100%) followed by parcinopy ($U_{95}^{\rho_{tot}} = 0.03203$ with a participation rate of 15.8%) that had the lowest values of $U_{x_*^m}$. Given that the bias for models librat, pbrt, rayspread, and RGM-2 (as well as inform in the NIR) was relatively constant (see section 5.1), it was considered inappropriate to compute $\tilde{f}_{95}^{\rho_{tot}}$ for these models. At the same time, it was found that the average value of $U_{95}^{\rho_{tot}}$ across all models was equal to 0.0235 in the red and 0.1561 in the NIR. These values amount to ~35% (~40%) of the average BRF values simulated in the red (NIR), i.e., 0.0676 (0.389), and were exceeded only by the models inform and RGM-2 (as well as RGM in the NIR only).

[67] The values of $U_{95}^{\rho_{tot}}$ reported in Table 4 are thus the de facto expanded uncertainties needed to explain 95% of a model's simulations in RAMI-IV. Correcting $U_{95}^{\rho_{tot}}$ by the coverage factor k yields the combined uncertainty of the simulations (u_c), which itself can be conceptualized as the result of two major contributions:

$$u_c = \sqrt{u_{c_{op}}^2 + u_{c_{mod}}^2} \qquad (3)$$

where $u_{c_{mod}}$ refers to combined standard uncertainty of the model, that is, the uncertainty contributions due to simplifications, parameterizations, and errors in the mathematical

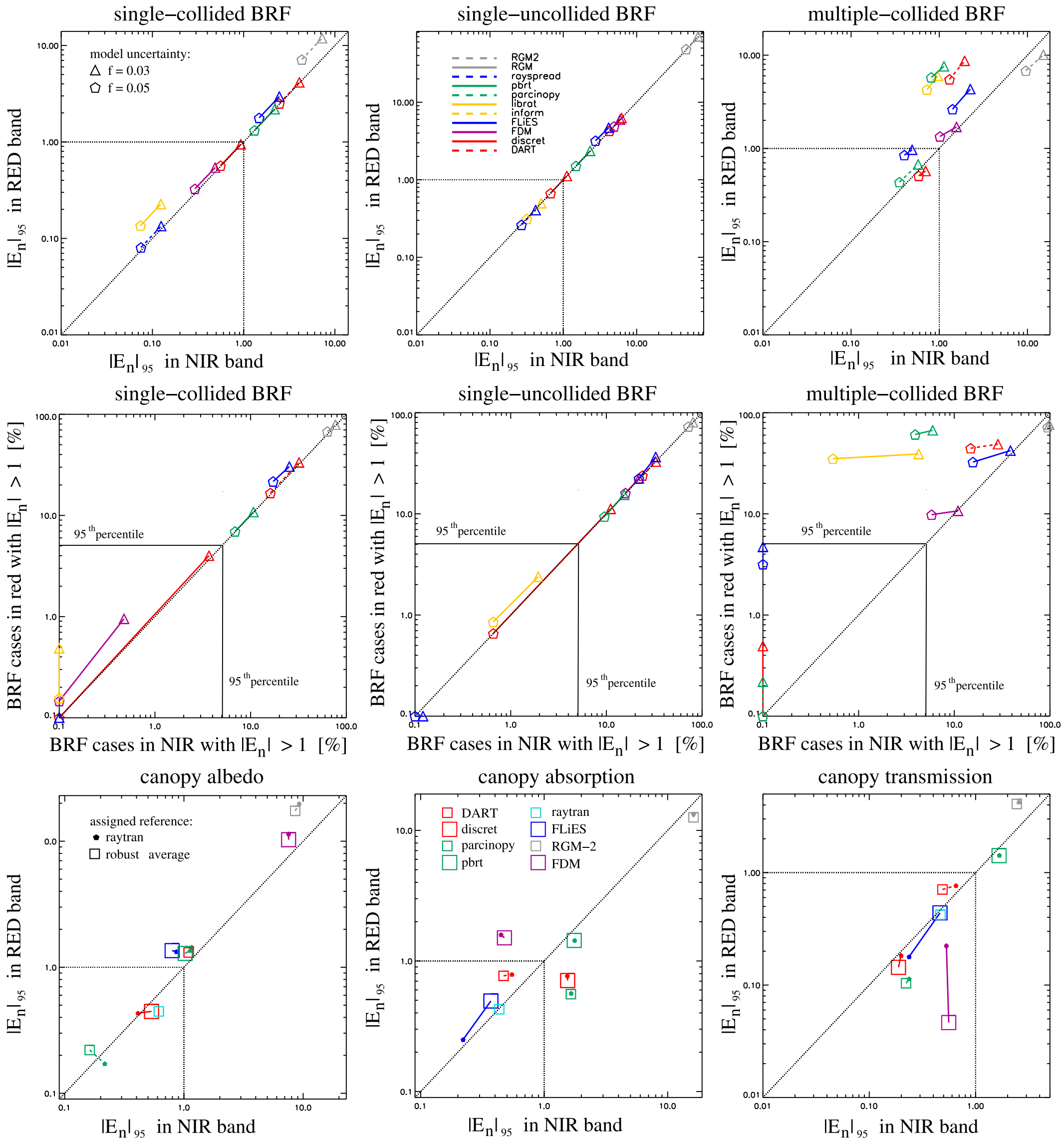

Figure 8. Graphs in the top and bottom rows show the differences between the red and NIR values of the 95th percentile generated from all $|E_n|$ numbers obtained for the "abstract canopy" test cases. When $|E_n|_{95} < 1$, the uncertainties associated with both the model and reference values are larger than their respective differences such that model and reference become de facto indiscernible (at least for 95% of the BRF/flux simulations). For BRFs, the standard uncertainty of the model was assumed to be either 3% (triangle) or 5% (pentagon) of the assigned reference. The middle row panels display the percentage of cases for which $|E_n| > 1$ in the red (or NIR): Models below (to the left of) the 95th percentile line for the red (NIR) have $|E_n| < 1$ in the corresponding top row panels. The bottom row shows $|E_n|_{95}$ for hemispherical flux simulations using two different assigned references values (symbols). The standard uncertainty associated with the flux simulations is derived from the corresponding GCOS accuracy criteria.

formulation of the RT model. The remaining term, $u_{c_{op}}$, relates to the combined standard uncertainty of the operator, that is, uncertainty contributions which are the consequence of operator choices and mistakes. The latter, for example, include the usage of square instead of circular foliage scatterer. $u_{c_{op}}$ may also be affected by wrongly specified illumination and measurement conditions, a reduction of the scene dimensions, and statistical representations of volumes that should have been modelled as a number of discrete objects. With the available model simulations, it is, however, not possible to deliver an estimate of $u_{c_{mod}}$ (and thus also $u_{c_{op}}$). Dedicated experiments, like those described in *ISO 5725-2* [1994] could be a means to derive $u^2_{c_{mod}}$ for some model types although in practice this may prove too time consuming to do.

5.3. z′ Scores

[68] The z' scores described in section 7.6 of ISO-13528 are a means to investigate whether a laboratory/model is capable of matching the proficiency criteria ($\hat{\sigma}$) if the stan-

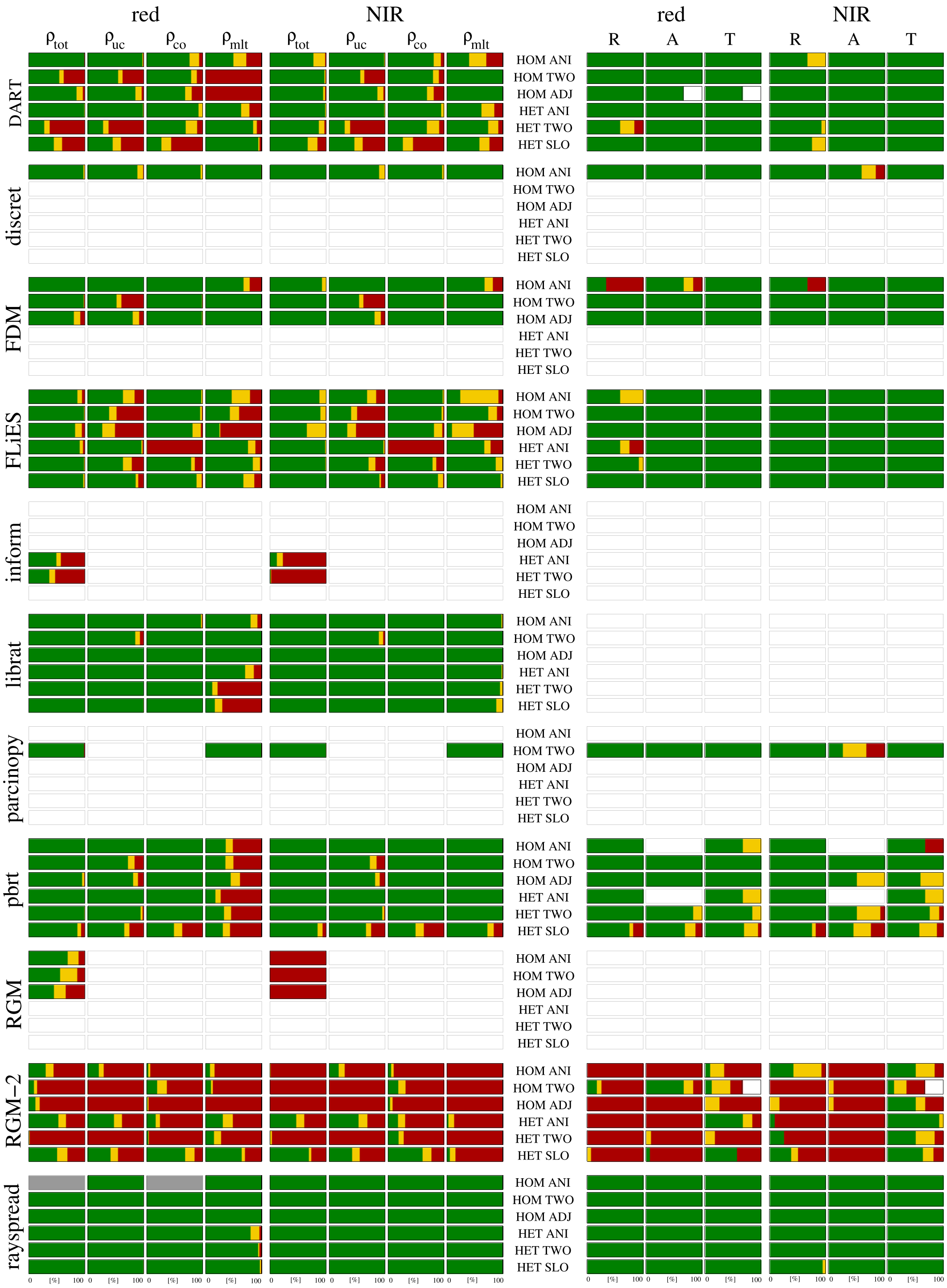

Figure 9. Bar charts showing different z' regimes for model simulated BRFs (left-hand graphs) and fluxes (right-hand graphs) in both the red and NIR. White indicates missing simulations, while grey indicates missing reference data. Shown are the fractions where $|z'| < 2$ (green), $2 \leq |z'| \leq 3$ (yellow, equivalent to a "warning signal" according to ISO-13528), and $|z'| > 3$ (red, equivalent to an "action signal" according to ISO-13528).

dard uncertainty of the assigned reference value (u_X) cannot be neglected. In RAMI-IV, the relevant metric is

$$z'(m;\lambda,\zeta,\Omega_v,\Omega_\mathrm{i}) = \frac{x_*^m(\lambda,\zeta,\Omega_v,\Omega_\mathrm{i}) - X_*(\lambda,\zeta,\Omega_v,\Omega_\mathrm{i})}{\sqrt{\hat{\sigma}^2(\lambda,\zeta,\Omega_v,\Omega_\mathrm{i}) + u_{X_*}^2(\lambda,\zeta,\Omega_v,\Omega_\mathrm{i})}} \quad (4)$$

where x_*^m is the radiative quantity of interest (e.g., a total BRF, a flux quantity, or one of their sub-components) simulated by model m for a given spectral (λ), structural (ζ), viewing (Ω_v), and illumination (Ω_i) related condition. X_* is the assigned model-specific reference value under these

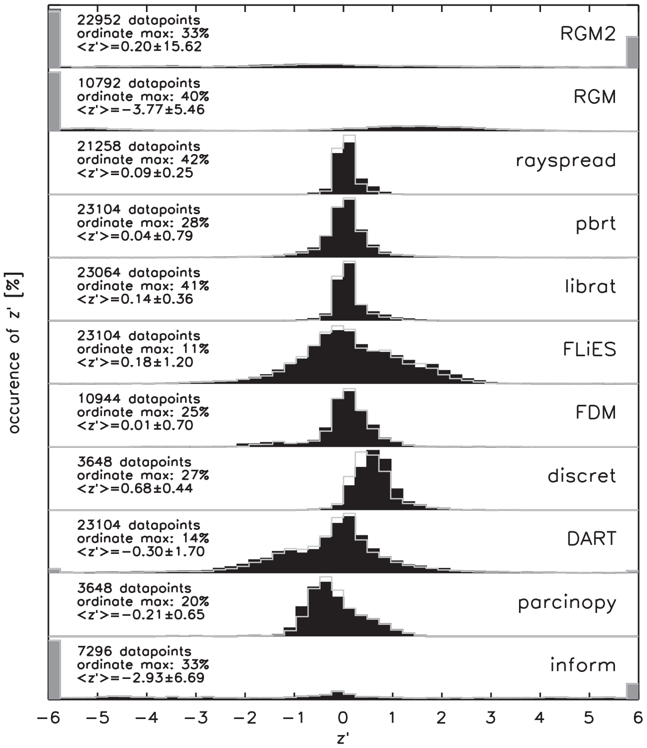

Figure 10. Histograms of z' statistics for total BRF simulations in RAMI-IV (abstract cases only). Histograms are scaled vertically, and z' counts outside the plot region are contained in the outermost grey bars. Black (filled) histograms used model-specific reference values ($X^m_{\rho_{tot}}$) to compute the z' scores, while the light grey colored histogram outline is based on a unique reference value for all models to compute the z' statistics ($X_{\rho_{tot}}$). Indicated in each plot are the number of simulated BRFs (maximum value = 23104), the maximum value of the vertical axis (ordinate max), as well as the mean and standard deviation of the z' distribution using the $X^m_{\rho_{tot}}$ approach.

conditions. The standard uncertainty of the reference u_{X_*} is defined in equation (A5) or at the end of sections 4.2.3 and 4.2.4. ISO-13528 instructs that z' scores above 2.0 or below –2.0 shall give rise to a "warning" sign, whereas those above 3.0 or below –3.0 should give rise to an "action" sign to the participant in question.

[69] Figure 9 provides detailed information on the occurrences of different z' regimes for model simulations of the various BRF components (left-hand panels) as well as the main hemispherical fluxes (right-hand panels) in both the red and NIR spectral domains. Shown are the percentages of simulations where $|z'| < 2$ (green), $2 \leq |z'| \leq 3$ (yellow), and $|z'| > 3$ (red). White indicates missing simulations, while grey indicates missing reference data. The standard uncertainty of the proficiency test was defined in item 3 of section 4.1 using $f = 0.03$ for the BRF simulations. Given the definition of E_n in section 5.2.1, it follows that $z' = 2E_n$, and hence, if the yellow and red parts of the bars of a given radiative quantity in Figure 9 are less or equal to 5% then that model should have $|E_n|_{95} \leq 1$ in the top or bottom rows of Figure 8.

[70] The first thing that one notices about Figure 9 is that half of the models did not deliver a full set of RAMI-IV simulations. This of course will have direct implications on the reliability of the proficiency statistics. Next, one will notice that whenever simulations were available, then the green colour generally dominates (in particular in the NIR). Model performance is very similar between the homogeneous and heterogeneous test cases with the exception perhaps of FLiES and RGM-2 which perform somewhat less well for the homogeneous test cases. None of the participating models, however, was capable of matching the GCOS accuracy criteria for surface albedo in 100% of the simulations in both the red and NIR.

[71] Some cases of "atypical" model behaviour—indicative of a possible dominance of $u_{c_{op}}$ in equation 3—can be identified in Figure 9. This concerns, for example, the DART model where one notices the unusually large number of action signs (red) for simulations of ρ_{mlt} over adjacent and two-layer homogeneous canopies in the red spectral band (and a complete absence of these in the NIR). Similarly, the FDM (FLiES) model results exhibit a different pattern for homogeneous (heterogeneous) test cases having anisotropic backgrounds. The model librat shows a spurious increase in action signs (red) for ρ_{mlt} simulations over the two-layer heterogeneous test cases as well as those having inclined tree crowns. For the pbrt model, the $|z'|$ statistics for the ρ_{uc} simulations of the inclined crown test cases are markedly different.

[72] If the distribution of biases is normal and both X_* and $\hat{\sigma}_*$ are good estimates of the mean and standard deviation of the population from which the simulated x^m_* are taken, then about 5% of the data will have $|z'| > 2$ and 0.3% can be expected of having $|z'| > 3$. Five percent in Figure 9 corresponds to the horizontal extent of the yellow segment in the third bar (counted from the right) of the bottom row. A separate analysis (not shown) showed that apart from the librat and rayspread models, only the discret model had less than 0.3% of its simulations above $z' = 3$ (both FDM and parcinopy were almost as good with $\sim$ 0.5% of the simulations having $z' > 3$).

[73] Figure 10 displays histograms of the z' statistics obtained from total BRF simulations over all abstract canopy test cases in RAMI-IV. More specifically, black-filled histograms used model-specific reference values ($X^m_{\rho_{tot}}$) to compute the z' scores, while the light grey colored histogram outline is based on the usage of a unique reference value across all models ($X_{\rho_{tot}}$) when computing the z' score. The latter approach tends to result in narrower and more peaked histograms because the target model is correlated with the reference (which is why ISO-13528 does not recommend it). While most histograms are mono-modal, their dispersion, skewness, and central locations vary considerably. The FDM model is the closest to having a zero mean z', but its skewness is –1.2 (instead of zero). Bimodal histograms occur for RGM and inform—and to a lesser extend also for DART—and are indicative of model performances that are substantially different between the red and NIR.

6. Concluding Remarks

[74] ISO-13528 was developed "to determine the performance of individual laboratories for specific tests or

measurements." This contribution made use of the ISO-13528 standard to evaluate physics-based computer simulation models that mimic the transfer of radiation in vegetation canopies. After a series of initial consistency checks, this procedure involved (1) the definition of a tolerance criteria suitable for the determination of proficiency in RT model simulations, (2) the definition of a sufficiently precise reference solution against which the candidate models could be compared, and (3) the selection of appropriate evaluation metrics to quantify the performance of the RT models.

[75] The choice of proficiency criteria ($\hat{\sigma}$) is crucial for the outcome of intercomparison efforts. For BRF simulations, we set $\hat{\sigma}$ to 3% and 5% of the reference value in analogy with vicarious calibration efforts. For hemispherical fluxes, we made use of the GCOS accuracy criteria for surface albedo and FAPAR to derive $\hat{\sigma}$. Due to missing information in *GCOS* [2011], we assumed a (type B) rectangular distribution of the biases within the maximum tolerable deviation ranges proposed by GCOS. This choice obviously has an impact on the likelihood of models to comply with the proficiency test. Ideally, GCOS should provide all necessary elements to allow for an unambiguous evaluation of ECV compliance. As a minimum, the percentiles of the ECV population that must comply with the GCOS accuracy criteria should be provided. For example, one could specify that 100%, 99%, or only 95% of the ECV retrievals (within a given time frame or geographical region) must fall within the GCOS recommended accuracy level. Additionally, the shape of the distribution of biases could be specified.

[76] The reliability of the reference solution is another important item in proficiency testing. For single-scattered BRF components, it was possible to use the simulations of the librat and rayspread Monte Carlo models because they matched the analytical solutions to within a fraction of 1%. For the multiple-scattered BRF component, the robust analysis method that is proposed in Annex C of ISO-13528 was used. For hemispherical fluxes, both the simulations of the raytran Monte Carlo model and the robust analysis approach were used. ISO-13528 points out that the standard uncertainty of the reference solution ideally should not exceed 30% of the proficiency criteria ($\hat{\sigma}$). This requirement, however, was rarely satisfied when the robust average approach was used to assign a reference value. Wherever possible, efforts should thus focus on identifying and using credible reference models instead.

[77] It was found that some of the more deviating models had never participated in previous phases of RAMI. Whether the differences observed in the current model comparison effort were caused by operator errors/choices or were genuine to the RT formulation/implementation of these models could not be determined on the basis of the available data. However, the fact that all test cases in RAMI-IV were new and that some "atypical" model behavior was noted in Figure 9, both point to an increased likelihood of operator-induced errors. In fact, RAMI-1 to RAMI-3 showed that the repetition of a given set of experiments in successive intercomparison rounds lead to a gradual improvement of (most) models. This is so because developers gradually identify and remove model weaknesses and software errors, and improve the manner in which the RAMI test cases are implemented in the participating models. A more rigorous approach to this matter would require evaluating the repeatability of Monte Carlo models as proposed in ISO 5725-2 and comparing the results with the actual expanded uncertainties as discussed in section 5.2.2.

[78] While the number of measurements that are used in laboratory intercomparisons is relatively small (typically < 100), the number of total BRF simulations requested for this proficiency test exceeded 20,000. It is thus not surprising that none of the participating RT models had a 100% success rate (i.e., $z' < 2$) for total BRF simulations. The only exception to this are the librat and rayspread models that were, however, used to assign the reference solution for the single-scattered BRF components. Similarly, for surface albedo simulations in the red and NIR, not a single participant was always matching the GCOS accuracy criteria for all six of the prescribed canopy architecture types. If the required success rate is reduced to 95% or even 90%, then the number of compliant models increases of course.

[79] Gaussian distributions of the biases and z' scores enable one to verify whether the number of outliers (i.e., the number of biases > $3\hat{\sigma}$ or the number of z' scores > 2 or 3) is actually in line with the theoretical expectation. Apart from model and operator biases, the non-Gaussian distributions of z' may also be due to inadequate sampling of canopy scenarios from the overall population of possible canopy architectures. In fact, if one ignores spectral, illumination, and viewing conditions, only 19 different canopy architectures remain (10 for the HOM cases and 9 for the HET). Some RT models submitted simulation results for less than a quarter of these. Such limited verification efforts possess little weight regarding the overall quality of a given model and furthermore complicate the interpretation of results from different models. The latter is because some operators apply their models only to test cases for which they were designed (e.g., 1-D models were not applied to 3-D test cases), while others run their model also on experiments that are likely to cause larger deviations (e.g., inform was designed for actual canopies but also delivered results for the abstract canopy cases).

[80] RAMI advocates a universal model verification strategy based on sufficiently large sets of test cases and qualified references. This is addressed through intercomparison rounds (RAMI phases) that are carried out at multi-annual intervals (i.e., 1999, 2002, and 2005). Each phase includes the experiments of the previous round (such as to enable participants to update/rectify their scores). Once reliable RAMI reference solutions have been established in this manner, they are transferred to the RAMI On-line Checker (ROMC) while RAMI continues with a new set of test cases. This approach is in line with ISO-13528 which specifically encourages continued proficiency testing. It also enables model developers (and users) to autonomously verify the quality of a given modelling tool via the ROMC instead of having to wait for the next phase of RAMI.

[81] Product certification is well suited to increase user confidence. In particular, in the manufacturing industries, new products can only enter the market after having passed a set of rigorous tests that certify their compliance with predefined quality objectives. By analogy, model certification would be a highly desirable aspect in efforts to improve the credibility of studies relying on canopy RT simulations. Conceptually, the development of a model certification platform requires access to (1) reliable procedures and criteria to

assess and quantify the performance of models and (2) easy-to-use interfaces allowing interested parties to autonomously perform model evalutions. The latter can be addressed by web-based facilities, like the ROMC. The former should be addressed by international standards, like ISO-13528, that make use of agreed-upon methodologies. All elements are thus in hand now to move toward a community-approved QA/QC system capable of standardised, user-friendly, and application-specific certifications of RT model quality.

Appendix A: Robust Analysis

[82] The term "robust average" and "robust standard deviation" should be understood to mean estimates of the population mean or the population standard deviation calculated using a robust algorithm [*ISO 13528*, 2005]. The methodology presented below has been transcribed from that presented in annex C.1 of the ISO13528:2005 standard in order to match the nomenclature and context of the RAMI exercise.

[83] If N models have generated a BRF value for a given viewing and illumination configuration, then denote the different values of BRFs, when sorted into increasing order, by $\rho_1, \rho_2, \ldots, \rho_i, \ldots, \rho_N$. Next denote the robust average and robust standard deviation of these data by ρ^* and s^*_ρ. Calculate initial values for ρ^* and s^*_ρ as

$$\rho^* = \text{median of } \rho_i \quad (i = 1, 2, \ldots, N) \tag{A1}$$

$$s^*_\rho = 1.483 \cdot \text{ median of } |\rho_i - \rho^*| \quad (i = 1, 2, \ldots, N) \tag{A2}$$

[84] Update the values of ρ^* and s^*_ρ as follows. First, compute $\delta = 1.5 \cdot s^*_\rho$. Then, for each ρ_i ($i = 1, 2, \ldots, N$), calculate

$$\rho^*_i = \begin{cases} \rho^* - \delta & \text{if } \rho_i < \rho^* - \delta \\ \rho^* + \delta & \text{if } \rho_i > \rho^* + \delta \\ \rho_i & \text{otherwise} \end{cases}$$

[85] Next, compute the new values of ρ^* and s^*_ρ from

$$\rho^* = \frac{1}{N}\sum_i^N \rho^*_i \tag{A3}$$

$$s^*_\rho = 1.134 \cdot \sqrt{\frac{1}{N-1}\sum_i^N \left(\rho^*_i - \rho^*\right)^2} \tag{A4}$$

[86] The robust estimates of ρ^* and s^*_ρ may then be derived by an iterative process, that is, by updating the values of ρ^* and s^*_ρ using the modified data until the process finally converges. Convergence was assumed to exist when ρ^* was stable within the precision requirements imposed by RAMI, i.e., to within 10^{-6}.

[87] Finally, the standard uncertainty u_X of the robust mean can be estimated as

$$u_X = 1.25 \cdot s^*_\rho/\sqrt{N} \tag{A5}$$

where according to *ISO 13528* [2005], the "factor 1.25 represents the ratio of the standard deviation of the median to the standard deviation of the arithmetic mean, for large samples ($N > 10$) from a normal distribution. For normally distributed data, the standard deviation of a robust average calculated using the algorithm in this appendix is not known, but will fall somewhere between the standard deviation of the arithmetic mean and the standard deviation of the median, so the above formula gives a conservative estimate of the standard uncertainty."

[88] **Acknowledgments.** RAMI would not have been possible without the support of Alan Belward, head of the Land Resource Management unit, and Frank Raes, head of the Climate Risk Monitoring unit, both at the Institute for Environment and Sustainability of the European Commission's Joint Research Centre in Ispra, Italy. The constructive comments of two anonymous reviewers and Jouni Peltoniemi on earlier drafts of this manuscript are gratefully acknowledged. The technical assistance of Monica Robustelli and the helpful advice of and discussions with the members of the RAMI Advisory Body is also thankfully acknowledged here.

References

Atzberger, C. (2000), Development of an invertible forest reflectance model: The infor-model, in *A Decade of Trans-European Remote Sensing Cooperation. Proceedings of the 20th EARSeL Symposium Dresden, Germany, 14–16 June 2000*, edited by M. Buchroithner, pp. 39–44, CRC Press/Balkema, Leiden, The Netherlands.

Bruegge, C. J., N. L. Chrien, R. R. Ando, D. J. Diner, W. A. Abdou, M. C. Helmlinger, S. H. Pilorz, and K. J. Thome (2002), Early validation of the multi-angle imaging spectroradiometer (MISR) radiometric scale, *IEEE Trans. Geosci. Remote Sens.*, *40*, 1477–1492.

Chelle, M. (1997), Développement d'un modèle de radiosité mixte pour simuler la distribution du rayonnement dans les couverts vegetaux, PhD Thesis, Université de Rennes I. ftp://ftp.irisa.fr/techreports/theses/1997/chelle.ps.gz.

Chelle, M. (2006), Could plant leaves be treated as Lambertian surfaces in dense crop canopies to estimate light absorption? *Ecol. Modell.*, *198*, 219–228.

Chen, J. M., and J. Cihlar (1995), Plant canopy gap size analysis theory for improving optical measurements of leaf area index, *Appl. Opt.*, *34*, 6211–6222.

Disney, M. I., P. Lewis, and P. R. J. North (2000), Monte Carlo raytracing in optical canopy reflectance modelling, *Remote Sens. Rev.*, *18*, 163–196.

Disney, M. I., P. Lewis, M. Bouvet, A. Prieto-Blanco, and S. Hancock (2009), Quantifying surface reflectivity for spaceborne lidar via two independent methods, *IEEE Trans. Geosci. Remote Sens.*, *47*(10), 3262–3271, doi:10.1109/TGRS.2009.2019268.

Gastellu-Etchegorry, J.-P., V. Demarez, V. Pinel, and F. Zagolski (1996), Modeling radiative transfer in heterogeneous 3-D vegetation canopies, *Remote Sens. Environ.*, *58*, 131–156.

Gastellu-Etchegorry, J.-P., E. Martin, and F. Gascon (2004), Dart: A 3D model for simulating satellite images and studying surface radiation budget, *Int. J. Remote Sens.*, *25*, 73–96.

GCOS (2011), Systematic observon requirements for satellite-based data products for climate (2011 update), supplemental details to the satellite-based component of the 'implementation plan for the global observing system for climate in support of the UNFCCC (2011 update)'. *GCOS-154*, Global Climate Observing System (GCOS), World Meteorogical Organisation.

Gerboles, M., et al. (2011), Interlaboratory comparison exercise for the determination of As, Cd, Ni and PM_{10} in Europe, *Atmos. Environ.*, *45*, 3488–3499.

Gobron, N., B. Pinty, M. M. Verstraete, and Y. Govaerts (1997), A semi-discrete model for the scattering of light by vegetation, *J. Geophys. Res.*, *102*, 9431–9446.

Goel, N. (1988), Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data, *Remote Sens. Rev.*, *4*, 1–212.

Goel, N. S., and D. E. Strebel (1984), Simple beta distribution representation of leaf orientation in vegetation canopies, *Agron. J.*, *76*, 800–803.

Govaerts, Y. (1995), A model of light scattering in three-dimensional plant canopies: A Monte Carlo ray tracing approach, Ph. D. thesis, Université Catholique de Louvain-la-Neuve, Département de Physique, 2, Chemin du Cyclotron, B–1348 Louvain-la-Neuve, Belgique.

Huang, H., M. Chen, and Q. Liu (2009), A realistic structure model for large-scale surface leaving radiance simulation of forest canopy and accuracy assessment, *Int. J. Remote Sens.*, *30*(20), 5421–5439.

Hund, E., D. L. Massart, and J. Smeyers-Verbeke (2000), Inter-laboratory studies in analytical chemistry, *Anal. Chim. Acta*, *423*, 145–165.

ISO 13528 (2005), Statistical methods for use in proficiency testing by interlaboratory comparisons, *International Standard, ISO*

13528:2005(E), 66 pp., International Organization for Standardization, ISO/TC6-SC6, Geneva, Switzerland.

ISO 5725-2 (1994), Accuracy (trueness and precision) of measurement methods and results—Part 2:, *International Standard, ISO 5725-2:1994 (E)*, 42 pp., International Organization for Standardization, ISO/TC69-SC6, Geneva, Switzerland.

ISO 5725-3 (1994), Accuracy (trueness and precision) of measurement methods and results—Part 3:, *International Standard, ISO 5725-3:1994 (E)*, 25 pp., International Organization for Standardization, ISO/TC69-SC6, Geneva, Switzerland.

JCGM (2008), Evaluation of measurement data—Guide to the expression of uncertainty in measurement, *GUM 1995 with minor corrections JCGM 100:2008*. Joint Committee for Guides in Metrology (JCGM), Bureau international des poids et mesures (BIPM) and other members of JCGM.

JCGM (2012), International vocabulary of metrology—Basic and general concepts and associated terms (VIM), 3rd edition, *VIM 2008 with minor corrections JCGM 200:2012*. Joint Committee for Guides in Metrology (JCGM), Bureau international des poids et mesures (BIPM) and other members of JCGM.

Kallel, A. (2010), Vegetation radiative transfer modeling based on virtual flux decomposition, *J. Quant. Spectrosc. Radiat. Transfer*, *111*, 1389–1405.

Kallel, A. (2012), Extension of virtual flux decomposition model to the case of two vegetation layers: FDM-2, *J. Quant. Spectrosc. Radiat. Transfer*, *113*, 440–460.

Kneubühler, M., M. Schaepman, K. Thome, F. Baret, and A. Müller (2002), Calibration and validation of Envisat MERIS. Part 1: Vicarious calibration at Rail Road valley Playa (NV), *Proceedings of MERIS level 2 validation Workshop, ESRIN*, Frascati, Italy, December 9–13.

Kobayashi, H., and H. Iwabuchi (2008), A coupled 1-D atmosphere and 3-D canopy radiative transfer model for canopy reflectance, light environment, and photosynthesis simulation in a heterogeneous landscape, *Remote Sens. Environ.*, *112*, 173–185.

Lewis, P. (1999), Three-dimensional plant modelling for remote sensing simulation studies using the botanical plant modelling system, *Agron. Agric. Environ.*, *19*, 185–210.

Liu, Q. H. H. H. G., W. Qin, K. Fu, and X. Li (2007), An extended 3-D radiosity graphics combined model for studying thermal-emission directionality of crop canopy, *IEEE Trans. Geosci. Remote Sens.*, *45*, 2900–2918.

Pharr, M., and G. Humphreys (2010), *Physically Based Rendering: From Theory to Implementation*, 1167 pp., Morgan Kaufmann, San Fransisco.

Pinty, B., et al. (2001), The radiation transfer model intercomparison (RAMI) exercise, *J. Geophys. Res.*, *106*, 11937–11956.

Pinty, B., et al. (2004), The radiation transfer model intercomparison (RAMI) exercise: Results from the second phase, *J. Geophys. Res.*, *109*, D06210, doi:10.1029/2004JD004252.

Qin, W., and S. A. W. Gerstl (2000), 3-D scene modeling of semi-desert vegetation cover and its radiation regime, *Remote Sens. Environ.*, *74*, 145–162.

Rahman, H., B. Pinty, and M. M. Verstraete (1993a), Coupled surface-atmosphere reflectance (CSAR) model. 2. Semiempirical surface model usable with NOAA Advanced Very High Resolution Radiometer data, *J. Geophys. Res.*, *98*, 20,791–20,801.

Rahman, H., M. M. Verstraete, and B. Pinty (1993b), Coupled surface-atmosphere reflectance (CSAR) model. 1. Model description and inversion on synthetic data, *J. Geophys. Res.*, *98*, 20779–20789.

Schlerf, M., and M. Atzberger (2006), Inversion of a forest reflectance model to estimate structural canopy variables from hyperspectral remote sensing data, *Remote Sens. Environ.*, *100*, 281–294.

Thome, K. J. (2001), Absolute radiometric calibration of Landsat 7 ETM+ using the reflectance-based method, *Remote Sens. Environ.*, *78*, 27–38.

Thome, K. J., K. Arai, S. Tsuchida, and S. F. Biggar (2008), Vicarious calibration of ASTER via the reflectance-based approach, *IEEE Trans. Geosci. Remote Sens.*, *46*, 3285–3295.

Wang, Y., J. Czapla-Myers, A. Lyapustin, K. Thome, and E. Dutton (2011), Aeronet-based surface reflectance validation network (ASRVN) data evaluation: Case study for railroad valley calibration site, *Remote Sens. Environ.*, *115*, 2710–2717, doi:10.1016/j.rse.2011.06.011.

Widlowski, J.-L., T. Lavergne, B. Pinty, M. M. Verstraete, and N. Gobron (2006), Rayspread: A virtual laboratory for rapid BRF simulations over 3-D plant canopies, in *Computational Methods in Transport*, edited by G. Frank, pp. 211–231, Lecture Notes in Computational Science and Engineering Series, 48, Springer Verlag, Berlin. ISBN–10 3–540–28, 122–3.

Widlowski, J.-L., M. Robustelli, M. Disney, J.-P. Gastellu-Etchegorry, T. Lavergne, P. Lewis, P. J. R. North, B. Pinty, R. Thompson, and M. M. Verstraete (2007a), The RAMI On-line Model Checker (ROMC): A web-based benchmarking facility for canopy reflectance models, *Remote Sens. Environ.*, *112*(3), 1144–1150, doi:10.1016/j.rse.2007.07.016.

Widlowski, J.-L., et al. (2007b), The third radiation transfer model intercomparison (RAMI) exercise: Documenting progress in canopy reflectance modelling, *J. Geophys. Res.*, *112*, D09111, doi:10.1029/2006JD007821.

Widlowski, J.-L., et al. (2011), RAMI4PILPS: An intercomparison of formulations for the partitioning of solar radiation in land surface models, *J. Geophys. Res.*, *116*, G02019, doi:10.1029/2010JG001511.